Linear A

Entropy, Epigraphy, and the Limits of Decipherment

A Mathematical Analysis

Silvano Salvador

Contents

Abstract **9**

1 The Problem and Its Mathematical Structure **11**
1.1 The Inherited Debate and Its Epistemological Deficits 11
1.1.1 The Four Evidentiary Gaps 12
1.1.2 The Hypothesis Space $\mathcal{H}_{LA}$ 13
1.2 The Core Mathematical Object: A Formal Description of the Linear A Corpus 14
1.3 Formalizing the Hypothesis Comparison: MDL and Bayes Factors 15
1.4 Robustness Checks: Sensitivity of the Hypothesis Space to Corpus Selection . 16
1.5 Relation to the Companion Volumes on the Voynich Manuscript and Proto-Elamite 18
1.6 Structure of the Volume 19

2 Epigraphic and Stratigraphic Chronology: A Bayesian Dating Model **20**
2.1 The Inscribed Corpus and Its Find-Contexts 20
2.1.1 The Bayesian Chronological Model: Full Specification 21
2.2 The Destruction Horizon and the Administrative Tablet Cluster 22
2.3 Votive and Libation Inscriptions: A Longer Use-Life 23
2.4 Sign-Form Seriation Across the Chronological Range 24
2.4.1 Morphometric Features and the Seriation Regression 25
2.5 Scribal Hand Identification and Its Chronological Implications 26
2.6 Cross-Validating the Chronological Model Against Independent Anchors . . . 27

3 Entropy, Zipf's Law, and the Generative Process **28**
3.1 Shannon Entropy of the Linear A Syllabogram Sequence 28
3.2 Zipf's Law and the Word-Frequency Distribution 29
3.2.1 Zipf–Mandelbrot Fit and Its Interpretation 30
3.3 The Positional Entropy Profile and Its Diagnostic Value 31
3.3.1 The Sign Inventory and Corpus Statistics 33
3.3.2 Zeroth-Order Entropy and Its Interpretation 34
3.3.3 The Positional Entropy Profile in Full 35
3.3.4 The Quantitative Basis for the Sign Inventory Overlap 35
3.3.5 Transition Matrices: Structure and Sparsity 36
3.4 Autocorrelation and Long-Range Structure 36
3.5 Word-Length Distribution and Morphological Inference 38
3.6 Sign-Inventory Overlap and the Limits of Value Continuity 39

4 n-gram Models, Transition Matrices, and Cross-Entropy Comparisons **40**
4.1 Syllable Transition Matrices Under Linear-B-Derived Readings 40
4.2 Cross-Entropy Against Mycenaean Greek 41
4.2.1 The Cross-Entropy Gap and Its Significance 43

4.3 A Worked Cross-Entropy Computation, Step by Step 43
4.4 Cross-Entropy Against a Typological Reference Set 45
4.4.1 Individual Reference Language Cross-Entropies 47
4.5 Higher-Order Models and the Curse of Sparsity 48
4.5.1 Trigram Analysis: Qualitative Observations Under Sparsity 49
4.6 Mutual Information Between Sign Position and Genre 49
4.7 Smoothing, Sparsity, and the Reliability Frontier 50

5 The Reference Corpus and Formal Language Classification 52
5.1 Constructing a Typologically Diverse Reference Set 52
5.2 Constructing a Typologically Diverse Reference Set 52
5.3 Constructing a Typologically Diverse Reference Set 52
5.3.1 Reference Corpus Preparation and Tokenization 53
5.4 Formal Criteria for Language-Family Discrimination 54
5.5 Where Linear A Falls in the Classification Space 56
5.5.1 Principal Component Analysis of the Classification Space 57
5.6 Geographic and Typological Diversity of the Reference Set 58
5.7 Validating the Reference Corpus Against Held-Out Languages 59

6 The Logographic, Numerical, and Sigillary Subsystems 61
6.1 The Metrological System: A Self-Contained Formal Structure 61
6.2 Logogram Semantics and the Limits of Iconic Decipherment 62
6.3 Sealings, Roundels, and the Semiotics of Administrative Authentication . . . 63
6.4 Weight Systems and Cross-Site Metrological Standardization 64
6.5 Commodity Vocabulary and Its Cross-Site Distribution 65

7 Evaluating the Principal Language-Family Hypotheses 68
7.1 H1: The Hellenic Hypothesis (Formally Excluded) 68
7.2 H2: The Pre-Greek Anatolian Hypothesis 69
7.2.1 Formal Exclusion of H1: The Projection-Distance Evidence 70
7.2.2 The Pre-Greek Anatolian Hypothesis: A Detailed Statistical Assessment 70
7.3 H3: The Northwest Semitic Hypothesis 71
7.4 H4: The Language-Isolate Hypothesis 72
7.5 A Fifth Hypothesis: Pre-Indo-European Mediterranean Substrate 74

8 Bayesian Model Comparison and the MDL Synthesis 76
8.1 Prior Elicitation for the Four Language-Family Hypotheses 76
8.2 Prior Elicitation for the Four Language-Family Hypotheses 76
8.3 Prior Elicitation for the Four Language-Family Hypotheses 76
8.4 Bayes Factor Computation Across the Diagnostic Battery 77
8.4.1 Individual Bayes Factors from the Diagnostic Battery 78
8.5 The MDL Differential and Its Interpretation 79
8.5.1 MDL Scores and Their Computation 80
8.6 Synthesis: A Provisional Posterior Ranking 80
8.7 Sensitivity of the Posterior Ranking to Prior Specification 81

9 The Linear B Comparandum: Constraints from a Solved Sibling System 83
9.1 Ventris's Decipherment and Its Methodological Legacy 83
9.2 Ventris's Decipherment and Its Methodological Legacy 83
9.3 Ventris's Decipherment and Its Methodological Legacy 83
9.3.1 The Grid Method and Its Transfer to the Linear A Problem 84
9.4 Sign-Value Transfer and Its Formal Limits 85

© Silvano Salvador

9.4.1 The Value-Continuity Assumption: Evidence For and Against 87
9.5 Quantifying the Information Content of the Linear B Comparandum 87
9.6 Why a Solved Sibling Script Has Not Solved Linear A 88
9.7 Computational Decipherment Attempts: A Formal Post-Mortem 90

10 Linear A in Comparative Perspective: Other Undeciphered Scripts 92
10.1 Cretan Hieroglyphic: The Possible Ancestor Script 92
10.2 Cypro-Minoan: A Structurally Distinct Sibling Problem 93
10.3 The Phaistos Disc: A Cautionary Case in Corpus Size 94
10.4 Linear Elamite: A Decipherment Achieved, and Its Lessons 95
10.5 A Cross-Script Summary Table 96

11 A Historiography of Linear A Scholarship 97
11.1 The Evans Era and the Initial Cataloguing Phase (1900–1952) 97
11.2 The Ventris Decipherment and Its Immediate Aftermath (1952–1980) 97
11.3 The Corpus-Consolidation Phase: GORILA and Its Successors (1976–2000) . 98
11.4 The Digital and Statistical Turn (2000–Present) 99
11.5 Quantitative Trends in Linear A Publication Volume 100

12 Network and Graph-Theoretic Analysis of the Sealing Administration 102
12.1 Sealings as a Relational Data Structure 102
12.2 Community Detection and Site-Level Modularity 103
12.3 Random-Graph Null Models and Statistical Significance 103
12.4 Word-Token Co-occurrence and Latent Role Categories 104
12.5 Comparison with the Linear B Sealing Network 104

13 Robustness, Simulation, and Sensitivity Analysis 106
13.1 Motivation: Quantifying Uncertainty in a Small, Damaged Corpus 106
13.1.1 Sources of Uncertainty 106
13.2 Bootstrap Resampling of Entropy Estimates 107
13.3 Sensitivity of the Hellenic-Exclusion Result 107
13.4 Simulation Study: Statistical Power Against Alternative Generative Models . . 108
13.5 Sensitivity of the Bayesian Model Comparison to Prior Specification 109

14 Missing-Data Imputation and Sign-Frequency Forecasting 110
14.1 The Problem of Damaged and Fragmentary Tablets 110
14.2 A Hidden Markov Model for Sign Sequence Imputation 110
14.3 Terminal Lacunae and Extrapolative Forecasting 111
14.4 Total Lacunae: Genre and Format Inference Without Sign-Level Recovery . . 111
14.5 Forecasting Future Discoveries: A Capture–Recapture Estimate 112

15 Toponymic and Onomastic Evidence Across the Aegean Bronze Age 113
15.1 Place-Names and Personal Names as a Distinct Evidentiary Category 113
15.2 Phonological Correspondence with Later Cretan Toponymy 113
15.3 Personal Names and the Reduplication Problem 114
15.4 Comparative Onomastics: Anatolian, Semitic, and Pre-Greek Strata 114

16 Quantitative Cross-Script Benchmarking: A Wider World Sample 116
16.1 Beyond the Linear B Comparandum: A Twelve-Script Benchmark 116
16.2 A Two-Dimensional Information-Theoretic Map 116
16.3 Robustness Across Corpus-Size Normalization 117
16.4 Implications for the Linear A Natural-Language Hypothesis 118

© Silvano Salvador

17 The Economics of Scribal Labor: Workload, Specialization, and Output 119
17.1 Tablets as Evidence of Labor Organization . . . 119
17.2 Output Volume and Specialization by Hand . . . 119
17.3 Estimating Training Time from Sign-Formation Consistency . . . 120
17.4 An Estimate of Total Scribal Population . . . 120

18 Metrological Systems: A Formal Reconstruction of Weights and Measures 122
18.1 Numerical Notation as a Tractable Sub-System . . . 122
18.2 Recovering Fractional Values from Summation Constraints . . . 122
18.3 Cross-Commodity Consistency of the Recovered System . . . 123
18.4 Implications for Administrative Capacity . . . 123

19 Quantitative Iconography of the Sealing Corpus 125
19.1 Seal Devices as a Distinct, Non-Textual Evidentiary Channel . . . 125
19.2 Co-occurrence Structure Between Motif Category and Administrative Role . . . 125
19.3 Stylistic Seriation of Motif Execution . . . 126

20 Trade Networks and Foreign Contact: Epigraphic Evidence from Abroad 127
20.1 Linear A Outside Crete: The Extra-Cretan Corpus . . . 127
20.2 Geographic Distribution and the Maritime Network . . . 127
20.3 Implications of Class (ii) Objects for the Language Question . . . 128
20.4 Sign-Inventory Divergence at Extra-Cretan Sites . . . 128

21 Formal Models of Scribal Error and Orthographic Variation 129
21.1 Distinguishing Genuine Variation from Scribal Error . . . 129
21.2 Error Rate by Scribal Hand and Its Correlates . . . 129
21.3 Re-estimating Entropy After Error Correction . . . 130

22 A Decision-Theoretic Framework for Evaluating Future Decipherment Claims 131
22.1 The Recurring Problem of Unfalsifiable Decipherment Announcements . . . 131
22.2 Relationship to the Monograph's Own Methodology . . . 132

23 Spatial Statistics of Tablet Find-Contexts Within Sites 134
23.1 Intra-Site Spatial Distribution as an Evidentiary Source . . . 134
23.2 Ripley's K-Function Analysis of Genre Clustering . . . 134
23.3 Cross-Site Comparison of Spatial Organization . . . 135

24 A Formal Treatment of the Word-Divider System 136
24.1 The Word-Divider as an Independently Informative Sign Class . . . 136
24.2 Comparison Against Word-Length Distributions in Known Languages . . . 136
24.3 Inconsistent Divider Usage and What It Reveals . . . 137
24.4 The Secondary Punctuation Mark . . . 137

25 Climate, Destruction Horizons, and the End of the Linear A Tradition 138
25.1 The Chronological Endpoint of the Corpus . . . 138
25.2 Tablet Corpus Concentration at the Destruction Horizon . . . 138
25.3 The Theran Eruption: Chronological Relationship to the Corpus . . . 139
25.4 The End of the Tradition and Its Replacement . . . 139

© Silvano Salvador

26 Material Analysis of Inscribed Surfaces and Its Evidentiary Value 141
26.1 Clay Fabric as an Independent Provenance Indicator . . . 141
26.2 Fabric Group Distribution Across Sites and Genres . . . 141
26.3 Pigment Traces and Their Limited Evidentiary Value . . . 142

27 A Comparative Survey of Decipherment Methodologies Across Scripts 143
27.1 What Successful Decipherments Have in Common . . . 143
27.2 The Bilingual-Text Criterion and Its Absence for Linear A . . . 143
27.3 Why Linear B's Success Has Not Transferred Cleanly to Linear A . . . 144

28 Pedagogical Appendix: Worked Statistical Derivations for Non-Specialists 146
28.1 Purpose and Scope of This Appendix . . . 146
28.1.1 Shannon Entropy from First Principles . . . 146
28.1.2 The Miller–Madow Bias Correction . . . 147
28.1.3 Zipf's Law and Log-Log Regression . . . 147
28.1.4 Bayes Factors and the MDL Criterion . . . 147

29 Computational Stemmatics: Scribal Schools and the Transmission of Sign-Form Variants 149
29.1 From Individual Hands to Scribal Lineages . . . 149
29.2 Hierarchical Clustering and the School Hypothesis . . . 150
29.3 A Network Model of Sign-Form Transmission . . . 150
29.4 Cross-Site Convergence and the Limits of the School Model . . . 151

30 The Libation Formula: A Formal Grammar of Religious Inscriptions 152
30.1 The Libation Formula as a Distinct Textual Genre . . . 152
30.2 Positional Slot Structure . . . 152
30.3 The Numerical Slot and Offering Quantification . . . 153
30.4 Geographic and Chronological Distribution of the Formula . . . 153

31 Authenticity, Provenance, and Forgery Detection in the Linear A Corpus 155
31.1 Why Authenticity Verification Matters for a Statistical Monograph . . . 155
31.2 Statistical Screening Against the Controlled-Excavation Baseline . . . 155
31.3 Sensitivity of Monograph Findings to the Unprovenanced Minority . . . 156

32 A Quantitative Theory of Sign Obsolescence and Innovation 158
32.1 The Life Cycle of a Script Sign . . . 158
32.2 Three Empirical Sign Trajectories . . . 158
32.3 Sign Innovation and the Rising Cluster . . . 159
32.4 Quantifying Sign Entropy Change Through Time . . . 160

33 Diachronic Entropy Drift and the Late Minoan Scribal Crisis 161
33.1 The LM IB Destruction Event and Its Administrative Aftermath . . . 161
33.2 Changes in Scribal Output Metrics Across the Crisis . . . 161
33.3 Entropy and Information Density Across the Crisis . . . 162
33.4 The Linear A–Linear B Transition as a System-Level Replacement . . . 163

34 Typological Universals and Linear A: A Cross-Linguistic Test 164
34.1 What Typological Universals Predict for Writing Systems . . . 164
34.2 Zipf's Law: A Universal Already Confirmed . . . 164
34.3 The Short-Form Preference Universal . . . 165
34.4 The Vocabulary Growth Universal . . . 165

© Silvano Salvador

34.5 The Non-Random Bigram Universal 166

35 Coda: What This Monograph Can and Cannot Establish 167
35.1 A Summary of the Positive Findings 167
35.2 The Principal Negative Findings 168
35.3 The Principal Unknowns 168
35.4 The Epistemic Standard Applied Throughout 169

36 A Formal Theory of Scribal Memory: Recurrence, Formulaicity, and Long-Range Dependence 171
36.1 The Concept of Scribal Memory 171
36.2 Burstiness and the Distinction Between Formulaic and Functional Words . . . 172
36.3 Long-Range Dependence in the Token Sequence 172
36.4 Formulaicity Index and Its Cross-Site Distribution 173

37 Quantitative Prosody: Rhythmic and Metric Constraints in the Linear A Textual Record 174
37.1 Can Prosodic Structure Be Recovered from an Undeciphered Script? 174
37.2 Word-Length Distribution 175
37.3 Word-Length Periodicity 175
37.4 The Sign-Position Entropy Gradient 176

38 Colophons, Catchwords, and Tablet-Series Linkage: Evidence for Multi-Tablet Dossiers 178
38.1 Do Linear A Tablets Form Series? 178
38.2 Detected Dossier Clusters at Ayia Triada 179
38.3 Implications for Corpus-Level Statistics 179

39 A Formal Model of the Linear A Administrative Year: Seasonal Cycles and Accounting Periodicity 181
39.1 Does the Linear A Archive Record a Seasonal Cycle? 181
39.2 Commodity-Type Clustering in Stratigraphic Sequence 181
39.3 Oil and Wine: A Third Accounting Season? 182

40 Synthesis and Conclusions 184
40.1 Summary of Findings Across the Four Axes 184
40.1.1 The Natural-Language Conclusion: A Consolidated Statement 185
40.1.2 The Language-Identification Problem: State of Play 185
40.2 Minoan as a Genuine, Unidentified Language 186
40.3 What Would Change the Posterior: A Pre-Registered Evidentiary Roadmap . 187
40.4 Open Problems and Directions for Further Research 188

A Formal Notation and Symbol Glossary 190
A.1 Corpus and Index Notation 191
A.2 Information-Theoretic and Statistical Notation 192

B Statistical Methods and Computational Details 193
B.1 Entropy Estimators and Bias Correction 193
B.2 Bootstrap Confidence Intervals 193
B.3 Permutation Tests 193
B.4 Corpus-Size Matching Experiment 194
B.5 Kneser–Ney Discount Parameter Estimation 194
B.6 Mahalanobis Distance Computation 194

© Silvano Salvador

B.6.1 Priority Research Directions 194
B.7 The MCMC Sampler for the Chronological Model 196
B.8 Pseudocode for the MDL/NML Score Computation 196
B.9 Handling Missing and Damaged Sign Data 197
B.10 Software, Data Status, and Reproducibility 198
B.11 Data Availability Statement 198

C Corpus Catalogue: Inscriptions by Site and Genre **200**
C.1 Cretan Find-Sites 200
C.2 Extra-Cretan Find-Sites 201
C.3 Genre Definitions and Coding Criteria 201

References **203**

© Silvano Salvador

Abstract

This monograph develops a mathematically rigorous framework for the analysis of Linear A, the undeciphered script of Bronze Age Crete, attested on roughly 1,400 inscribed objects — clay tablets, roundels, sealings, libation-table inscriptions, and votive offerings — spanning the Middle and Late Minoan periods (*c.* 1800–1450 BCE). Linear A occupies a position without parallel among the world's undeciphered scripts: it stands in close graphical kinship with Linear B, a structurally related but chronologically later script that *was* successfully deciphered by Michael Ventris in 1952 as an early form of Greek, and whose sign values, when mechanically applied to Linear A texts, yield sequences that are phonologically well-formed yet linguistically uninterpretable as any attested language — the single most important constraint in the entire literature. The work proceeds on four analytically independent axes, directly parallel to the Voynich framework but adapted to the unique evidentiary situation created by a partially-readable sibling script. The first is epigraphic and chronological: stratigraphic and ceramic-sequence dating of inscribed find-contexts across Cretan and extra-Cretan sites is combined with sign-form seriation to produce a formal chronology of the corpus. The second axis is information-theoretic: entropy estimation, Zipf-law regression, and n-gram divergence measurement are applied to the Linear A sign and syllabogram sequences, with systematic comparison against the much larger, fully transliterated Linear B corpus. The third axis is statistical-linguistic: conditional entropy profiles and transition-probability matrices, computed under the Linear-B-derived phonetic readings where applicable, are used to formally exclude candidate language families and to constrain the residual hypothesis space. The fourth axis is Bayesian model comparison: Bayes factors and Minimum Description Length differentials are computed across the principal competing hypotheses regarding the underlying "Minoan" language to produce a formally grounded posterior ranking.

The principal findings are as follows. The chronological model situates the bulk of the administrative tablet corpus within a narrow window in the Late Minoan IB period, consistent with the palatial destruction horizon, while votive and libation inscriptions show a substantially longer use-life extending into earlier Middle Minoan contexts. Applying Linear B phonetic values to Linear A sign sequences produces a syllable-transition-probability profile that is statistically incompatible, at significance levels exceeding $p < 0.001$, with Mycenaean Greek and with every other well-attested ancient language tested, formally excluding the Hellenic hypothesis under the assumption of value continuity between the two scripts. The conditional entropy profile of the Linear A sign sequence is broadly consistent with natural-language encoding, distinguishing it sharply from the more anomalous Voynich and proto-Elamite profiles, and supporting the conclusion that Linear A almost certainly encodes a genuine, as yet unidentified spoken language — provisionally labeled "Minoan" — rather than a restricted notation. The monograph concludes with a formal model comparison among the Pre-Greek Anatolian, Northwest Semitic, language-isolate, and residual Indo-European (non-Hellenic) hypotheses, finding that the language-isolate hypothesis currently achieves the lowest MDL score but by a margin that does not yet permit confident exclusion of the Anatolian alternative.

Keywords: Linear A, Linear B, Minoan Crete, information theory, conditional entropy, Zipf's law, Bayesian epigraphic chronology, n-gram analysis, MDL model comparison, undeciphered scripts, Michael Ventris.

Note on the Status of the Quantitative Material in This Monograph

This monograph is a *methodological* work demonstrating a mathematical and computational framework for evaluating decipherment and language-classification hypotheses, applied to Linear A as a case study. Readers should distinguish two categories of claim made in what follows.

Genuinely published, independently verifiable facts, cited and used as such, include: the existence and general scope of the GORILA corpus catalogue of Godart and Olivier; the decipherment of the structurally related Linear B script by Michael Ventris (1952) as an early form of Greek; the well-documented observation that mechanically applying Linear B phonetic values to Linear A sign sequences produces output that is phonologically well-formed but not interpretable as any attested language; and the existence of John Younger's ongoing digital extension of the Linear A corpus.

Illustrative, synthetic, or invented quantitative material, by contrast, comprises the great majority of the specific numbers reported in the analytical chapters: entropy values, Zipf exponents, Bayes factors, MDL differentials, bootstrap and permutation p-values, the morphometric scribal-hand clustering and network-centrality results, and most of the precise summary statistics attached to named Findings and Propositions throughout. No verified, fully tokenized digital transcription of the complete corpus at the precision implied by figures such as $N_{LA} = 7{,}362$ tokens, and no curated high-resolution photographic dataset sufficient for the morphometric stemmatic analysis of later chapters, was available to or independently verified by the author at the stated precision. These results are computed on, or illustrate the application of the method to, a synthetic dataset constructed to have plausible statistical properties, and should be read as worked demonstrations of *how* the framework would be applied to verified primary data, not as empirical findings about the real script or about any specific language hypothesis. Full disclosure of data status is given in the chapter *Statistical Methods and Computational Details* and its Data Availability Statement.

© Silvano Salvador

Chapter 1

The Problem and Its Mathematical Structure

1.1 The Inherited Debate and Its Epistemological Deficits

The Linear A problem is exceptional in the study of undeciphered scripts precisely because it is bounded on both sides by well-attested, well-studied systems. On one side lies the Cretan Hieroglyphic script, an earlier Minoan writing system whose relationship to Linear A is debated but whose geographical and cultural origin is not; on the other side lies Linear B, deciphered by Ventris in 1952 as an early form of Mycenaean Greek and sharing a substantial portion of its sign inventory with Linear A. This double bounding is without parallel among the world's undeciphered scripts: the Egyptian hieroglyphic decipherment relied on the Rosetta Stone's bilingual text, and the Mayan decipherment exploited structural regularities and a partially continuous scribal tradition. Linear A has no Rosetta Stone, but it does have something almost as tantalizing: a descendant (or sibling) script in which forty or more of its signs can be read phonetically, and in which those readings, mechanically applied, yield phonologically plausible but linguistically opaque sequences. This fact—that Linear A is partially *readable* but wholly *uninterpretable*—defines the epistemological frontier this monograph inhabits.

The quantitative literature on Linear A is thin compared to the Voynich corpus, and what exists has suffered from the same methodological deficits documented throughout that literature: informal claims of identification advanced without reproducible statistical tests, cherry-picked parallels substituting for systematic corpus analysis, and the absence of explicit null models against which decipherment candidates could be evaluated. The most influential quantitative contributions—John Younger's epigraphic catalogue, the GORILA corpus of Godart and Olivier, and more recently the digital extensions of those corpora by Finkelberg, Facchetti, and Tsikritsis—have been primarily descriptive rather than inferential. This monograph attempts to supply the missing inferential layer.

Definition 1.1: The Hypothesis Space $\mathcal{H}_{LA}$

Let $\mathcal{H}_{LA}$ denote the space of competing hypotheses regarding the linguistic affiliation of the Minoan language encoded in Linear A. Four principal hypotheses are formally evaluated throughout this monograph:

- $\mathcal{H}_{\text{ISO}}$: Minoan is a *language isolate*—genetically unrelated to any presently attested language family. Under this hypothesis, the Linear-B-derived phonetic readings of shared signs are accepted, but the resulting phonological sequences belong to

no classifiable family.

- $\mathcal{H}_{\text{ANAT}}$: Minoan belongs to the *Anatolian branch* of Proto-Indo-European, or to a pre-Greek Aegean substrate language exhibiting typological convergence with Anatolian languages. This hypothesis is motivated by proposed cognates between the "Pre-Greek" substrate vocabulary of later Greek and reconstructed Minoan roots, and by the Luwian typological parallels proposed by Woudhuizen (2002) and Hooker (1990).
- $\mathcal{H}_{\text{SEM}}$: Minoan belongs to the *Northwest Semitic* family. This minority hypothesis, associated with Cyrus Gordon (1966) and more recently with Jan Best and Fred Woudhuizen (1988), draws on proposed lexical correspondences in the libation-formula inscriptions.
- $\mathcal{H}_{\text{IE}}$: Minoan belongs to the Indo-European family but to a branch other than Hellenic or Anatolian—perhaps an extinct branch otherwise unattested, or an early stratum of the family predating the conventionally reconstructed PIE community.

All four hypotheses accept the prior conclusion, established on entropy grounds in Chapter 3, that Linear A encodes a genuine natural language. The Hellenic hypothesis $\mathcal{H}_{\text{HEL}}$ (Minoan encodes Greek or a closely related dialect) is formally excluded in Chapter 4 and thereafter treated as a *refuted* hypothesis rather than a live alternative.

Note 1.1: What This Monograph Does Not Claim

No decipherment of Linear A is proposed or implied by the formal results of this monograph. The methodological aim is the characterization of the hypothesis space and the formal posterior ranking of surviving language-family hypotheses. A finding that $\mathcal{H}_{\text{ISO}}$ achieves the lowest MDL score and the highest posterior probability over the evidence battery is a finding about the *most parsimonious current explanation*, not a finding that Minoan has been identified as an isolate in any strong linguistic sense. Future discovery of a bilingual inscription or a sufficiently large parallel text corpus could alter this posterior ranking substantially; the Bayesian framework adopted here is designed to accommodate such revision explicitly.

1.1.1 The Four Evidentiary Gaps

The scholarly frustration with the Linear A problem, unusual in its intensity even by the standards of undeciphered-script literature, is traceable to four specific evidentiary gaps—each independently surmountable in principle, yet jointly constituting a constraint set that no available technique has so far broken through.

Definition 1.1: The Four Evidentiary Gaps in Linear A Research

(i) **No bilingual text.** Unlike the Rosetta Stone's role in Egyptian decipherment or the Hittite–Akkadian bilinguals at Boghazköy, no inscription juxtaposing Linear A text with a securely understood language has been identified. Proposed partial bilinguals (including certain sealings from Knossos showing both Linear A and Linear B elements) have not survived close philological scrutiny.

(ii) **Structural knowledge without lexical knowledge.** Ventris's decipherment of

© Silvano Salvador

Linear B supplies candidate phonetic values for approximately 60–70% of the Linear A syllabic inventory (Section 3.4) but, applied to Linear A texts, yields phonological strings recognizable as well-formed syllabic sequences yet uninterpretable as any attested language.

(iii) **Corpus sparsity relative to vocabulary complexity.** With $N_{LA} \approx 7{,}362$ syllabic tokens across a sign inventory of $|\Sigma_{\text{syll}}| \approx 60$ signs, the ratio of corpus to alphabet is too low to reliably estimate more than first-order sequential statistics (Chapter 4), leaving higher-order structure—which would be most discriminating for language-family identification—statistically inaccessible.

(iv) **Absence of a confirmed linguistic descendant.** If Eteocretan (Section 6.7) is a descendant of Minoan, it provides a potentially powerful comparandum—but Eteocretan's own corpus is so small ($N \approx 150$ tokens) that establishing the genetic relationship with the rigor this monograph demands is not yet feasible, creating a circularity: Eteocretan would help decipher Minoan, but Minoan would help interpret Eteocretan.

Note 1.1: Why This Monograph Cannot Offer a Decipherment

The four gaps of Definition 1.1 are not gaps this monograph closes. Its contribution is, more modestly, to show how much can be rigorously established about the Linear A language-identification problem in the presence of these gaps—and, equally importantly, to quantify how much cannot. The formal Bayesian synthesis of Chapter 7 does not produce a decipherment; it produces a calibrated probability distribution over an explicitly limited hypothesis space, and it is as honest about its uncertainty as about its positive conclusions.

1.1.2 The Hypothesis Space $\mathcal{H}_{LA}$

Throughout this monograph the language-family identification problem is cast as model comparison over an explicitly enumerated hypothesis space.

Definition 1.2: The Corpus $\mathcal{C}_{LA}$ and Sub-Corpora

The full Linear A corpus $\mathcal{C}_{LA}$ comprises all 1,427 inscribed objects in the GORILA catalogue and its digital extensions. For analytical purposes, five sub-corpora are distinguished:

- $\mathcal{C}_{\text{tab}}$: administrative tablets ($n = 215$ objects, $N_{\text{tab}} \approx 6{,}200$ syllabic tokens);
- $\mathcal{C}_{\text{lib}}$: libation-table inscriptions ($n = 31$, $N_{\text{lib}} \approx 1{,}162$ tokens);
- $\mathcal{C}_{\text{seal}}$: sealings and nodules ($n = 341$, predominantly brief);
- $\mathcal{C}_{\text{rnd}}$: roundels ($n = 108$);
- $\mathcal{C}_{\text{vot}}$: votive and miscellaneous ($n = 732$, predominantly single sign-groups).

The entropy and n-gram analyses of Chapters 3–5 operate on $\mathcal{C}_{\text{tab}} \cup \mathcal{C}_{\text{lib}}$ (total $N_{LA} \approx 7{,}362$ syllabic tokens) as the primary analytical corpus, since the sealing, roundel, and votive sub-corpora are either too brief per object or too administratively formulaic to contribute informative sequential structure.

© Silvano Salvador

Definition 1.3: MDL and the NML Criterion

Given data $\mathcal{D}$ and a parametric model class $\mathcal{M}_{mathcalH}$ encoding hypothesis $\mathcal{H}$, the Normalized Maximum Likelihood (NML) code length is

$$L_{\text{NML}}(\mathcal{D} \mid \mathcal{M}_{mathcalH}) = -\log \frac{p(\mathcal{D} \mid \hat{\theta}(\mathcal{D}))}{\sum_{\mathcal{D}'} p(\mathcal{D}' \mid \hat{\theta}(\mathcal{D}'))},$$

where $\hat{\theta}(\mathcal{D})$ is the maximum-likelihood estimator and the sum in the denominator (the *stochastic complexity* of $\mathcal{M}_{mathcalH}$) runs over all data sequences of the same length. The MDL model selection criterion chooses $\mathcal{H}^* = \arg\min_{mathcalH} L_{\text{NML}}(\mathcal{D} \mid \mathcal{M}_{mathcalH})$, and the *MDL differential* $\Delta L = L_{\text{NML}}(\mathcal{D} \mid \mathcal{M}_{mathcalH}) - L_{\text{NML}}(\mathcal{D} \mid \mathcal{M}_{\mathcal{H}^*})$ quantifies the cost of choosing $\mathcal{H}$ over the optimal hypothesis.

1.2 The Core Mathematical Object: A Formal Description of the Linear A Corpus

We define the Linear A corpus following the GORILA catalogue (Godart and Olivier, 1976–1985, 5 vols.) and its subsequent digital extensions, treating the corpus as a formal combinatorial object amenable to information-theoretic analysis.

Definition 1.2: The Linear A Corpus $\mathcal{C}_{LA}$

Let Σ_{LA} denote the Linear A sign inventory under the conventional GORILA sign numbering, partitioned into three functionally distinct subsets:

- $\Sigma_{\text{syll}} \subset \Sigma_{LA}$: syllabograms, signs representing CV or V syllabic values. The attested syllabogram inventory comprises approximately 60 signs; of these, approximately 44 have a high-confidence Linear-B comparandum allowing phonetic value assignment (following the sign-correspondence table of Hooker 1980 and its revisions through Packard 2015).
- $\Sigma_{\text{log}} \subset \Sigma_{LA}$: logograms, signs representing commodity or item classes, attested primarily in administrative contexts.
- $\Sigma_{\text{num}} \subset \Sigma_{LA}$: numerical and metrological signs, forming a partially deciphered subsystem used for quantitative recording.

The corpus $\mathcal{C}_{LA}$ is then the ordered set of sign-group sequences

$$\mathcal{C}_{LA} = \{(s_1^{(j)}, s_2^{(j)}, \ldots, s_{n_j}^{(j)})\}_{j=1}^{M},$$

where $M \approx 1{,}427$ is the total number of inscribed objects, $s_k^{(j)} \in \Sigma_{LA}$, and n_j is the sign-group length of the j-th inscription. The corpus is further stratified by *genre* into five sub-corpora:

(i) $\mathcal{C}_{\text{tab}}$: clay tablets (administrative); $|\mathcal{C}_{\text{tab}}| \approx 147$ objects.

(ii) $\mathcal{C}_{\text{rnd}}$: roundels; $|\mathcal{C}_{\text{rnd}}| \approx 108$ objects.

(iii) $\mathcal{C}_{\text{seal}}$: sealings and nodules; $|\mathcal{C}_{\text{seal}}| \approx 341$ objects.

© Silvano Salvador

(iv) $\mathcal{C}_{\text{lib}}$: libation tables; $|\mathcal{C}_{\text{lib}}| \approx 31$ objects.

(v) $\mathcal{C}_{\text{vot}}$: votive and miscellaneous; $|\mathcal{C}_{\text{vot}}| \approx 800$ objects.

The syllabic sub-corpus used for all entropy and n-gram analyses is restricted to $\mathcal{C}_{\text{tab}} \cup \mathcal{C}_{\text{lib}}$, the two genres with sufficiently dense syllabographic text sequences; logogram-dominated inscriptions are excluded from entropy computations unless otherwise noted.

Finding 1.1: Corpus Size and the Sparsity Constraint

The total syllabographic token count of the Linear A corpus, restricted to $\mathcal{C}_{\text{tab}} \cup \mathcal{C}_{\text{lib}}$, is approximately

$$N_{LA} \approx 7{,}362 \text{ sign tokens},$$

compared to $N_{LB} \approx 85{,}000$ sign tokens in the Linear B corpus (Ventris and Chadwick 1973; Palaima 2011 digital extension). The ratio $N_{LA}/N_{LB} \approx 0.087$ defines the *sparsity regime* that constrains all higher-order statistical analyses in this monograph: bigram statistics are reliable to order $O(N_{LA}^{1/2}) \approx 86$ expected counts per cell at the population level, but trigram and higher-order models require explicit regularization (Chapter 4, Section 4.4). This sparsity is the primary technical obstacle distinguishing the Linear A analysis from the Voynich analysis, where $N_{\text{VM}} \approx 37{,}000$ tokens provide a substantially larger sample.

Proposition 1.1: Sign-Type Distribution and the Hapax Rate

Let f_s denote the frequency of sign type $s \in \Sigma_{\text{syll}}$ across $\mathcal{C}_{\text{tab}} \cup \mathcal{C}_{\text{lib}}$. The hapax rate—the proportion of sign types attested exactly once—satisfies

$$\rho_{\text{hap}} = \frac{|\{s : f_s = 1\}|}{|\Sigma_{\text{syll}}|} \approx 0.18,$$

which is consistent with the hapax rates observed in small epigraphic corpora of attested Bronze Age syllabic scripts (Linear B: $\rho_{\text{hap}} \approx 0.04$; Cypriot Syllabary: $\rho_{\text{hap}} \approx 0.09$) after controlling for corpus size via the Good–Turing estimator. This finding is used in Chapter 3 to argue that the Linear A sign inventory is not artificially inflated by scribal allographic variation beyond what is observed in comparable attested corpora.

1.3 Formalizing the Hypothesis Comparison: MDL and Bayes Factors

The MDL framework adopted in this monograph treats each language-family hypothesis $\mathcal{H} \in \mathcal{H}_{LA}$ as a generative model $\mathcal{M}_{\mathcal{H}}$ for the observed sign-group sequence data $\mathcal{D}$, and evaluates hypotheses by their *code length* under the normalized maximum likelihood (NML) criterion.

© Silvano Salvador

Definition 1.3: MDL Score and the NML Criterion

For a parametric model class $\mathcal{M}_{\mathcal{H}}$ with parameter space $\Theta_{\mathcal{H}}$, the Normalized Maximum Likelihood (NML) code length for data $\mathcal{D}$ is

$$L_{\mathrm{NML}}(\mathcal{D} \mid \mathcal{M}_{\mathcal{H}}) = -\log \frac{p(\mathcal{D} \mid \hat{\theta}(\mathcal{D}), \mathcal{M}_{\mathcal{H}})}{\sum_{\mathcal{D}'} p(\mathcal{D}' \mid \hat{\theta}(\mathcal{D}'), \mathcal{M}_{\mathcal{H}})},$$

where $\hat{\theta}(\mathcal{D})$ is the maximum-likelihood parameter estimate under $\mathcal{M}_{\mathcal{H}}$ and the sum in the denominator runs over all data sequences $\mathcal{D}'$ of the same length as $\mathcal{D}$. The MDL score for hypothesis $\mathcal{H}$ given data $\mathcal{D}$ is defined as

$$\mathrm{MDL}(\mathcal{H} \mid \mathcal{D}) = L_{\mathrm{NML}}(\mathcal{D} \mid \mathcal{M}_{\mathcal{H}}).$$

The *MDL differential* between two hypotheses $\mathcal{H}$ and $\mathcal{H}'$ is

$$\Delta\mathrm{MDL}(\mathcal{H}, \mathcal{H}' \mid \mathcal{D}) = \mathrm{MDL}(\mathcal{H}' \mid \mathcal{D}) - \mathrm{MDL}(\mathcal{H} \mid \mathcal{D}),$$

with positive values indicating that $\mathcal{H}$ is the more parsimonious model.

Definition 1.4: Bayes Factor for Language-Family Discrimination

The Bayes factor $\mathcal{K}(\mathcal{H}, \mathcal{H}')$ comparing hypotheses $\mathcal{H}$ and $\mathcal{H}'$ given data $\mathcal{D}$ is

$$\mathcal{K}(\mathcal{H}, \mathcal{H}') = \frac{p(\mathcal{D} \mid \mathcal{H})}{p(\mathcal{D} \mid \mathcal{H}')} = \frac{\int_{\Theta_{\mathcal{H}}} p(\mathcal{D} \mid \theta)\pi(\theta \mid \mathcal{H})\,\mathrm{d}\theta}{\int_{\Theta_{\mathcal{H}'}} p(\mathcal{D} \mid \theta')\pi(\theta' \mid \mathcal{H}')\,\mathrm{d}\theta'},$$

where $\pi(\theta \mid \mathcal{H})$ is the prior over the parameter space of $\mathcal{M}_{\mathcal{H}}$. The Jeffreys (1961) interpretation scale is used throughout: $\log_{10}\mathcal{K} \in [0, \frac{1}{2})$ is "not worth more than a bare mention"; $[\frac{1}{2}, 1)$ is "substantial"; $[1, 2)$ is "strong"; ≥ 2 is "decisive."

Note 1.2: The Relationship Between MDL and Bayes Factors

Under the NML parameterization, the MDL differential and the log Bayes factor are asymptotically equivalent for large corpora:

$$\Delta\mathrm{MDL}(\mathcal{H}, \mathcal{H}') \approx \log \mathcal{K}(\mathcal{H}, \mathcal{H}') + O\left(\frac{\dim(\Theta_{\mathcal{H}}) - \dim(\Theta_{\mathcal{H}'})}{2} \log N\right),$$

where the correction term is the standard BIC penalty for model complexity. In the small-corpus regime of Linear A ($N_{LA} \approx 7{,}362$), this approximation is imprecise enough that both quantities are computed and reported independently; conclusions that are robust to the choice of criterion are flagged explicitly.

1.4 Robustness Checks: Sensitivity of the Hypothesis Space to Corpus Selection

A natural objection to any quantitative treatment of Linear A is that the corpus itself is a contested object: editorial decisions in the GORILA catalogue regarding sign-group boundaries,

© Silvano Salvador

the treatment of damaged or partially legible signs, and the inclusion or exclusion of disputed inscriptions (the so-called "problematic" texts flagged by Godart and Olivier with an asterisk) could in principle drive the downstream statistical results. This section formalizes that concern as a sensitivity analysis and shows that the principal conclusions of the monograph are stable under plausible corpus perturbations.

Definition 1.5: Corpus Perturbation Family $\mathcal{P}_\delta$

Let $\mathcal{C}_{LA}$ denote the baseline corpus of Definition 1.2. For a perturbation parameter $\delta \in [0, 1]$, define the perturbed corpus family

$\mathcal{P}_\delta = \{\mathcal{C}' : \mathcal{C}'$ obtained from $\mathcal{C}_{LA}$ by removing a fraction δ of the most disputed inscriptions, ranked by ed

The editorial confidence score $c_j \in [0, 1]$ for inscription j is assigned following the GORILA damage-classification scheme: $c_j = 1$ for fully legible inscriptions, $c_j = 0.7$ for inscriptions with minor lacunae not affecting sign identification, $c_j = 0.4$ for inscriptions with disputed sign readings, and $c_j = 0.15$ for inscriptions flagged as "problematic" in Godart and Olivier (1976–1985). At perturbation level δ, the bottom δ-fraction of inscriptions by c_j are excluded.

Finding 1.2: Stability of Entropy Estimates Under Corpus Perturbation

Recomputing the unigram and bigram entropy estimates of Chapter 3 across the perturbation family $\mathcal{P}_\delta$ for $\delta \in \{0, 0.05, 0.10, 0.15, 0.20\}$ yields the following stability profile:

δ	N_{tab} retained	$\hat{H}_1$ [bits]	$\hat{H}_2$ [bits]	$\Delta\hat{H}_1$ from baseline
0.00	7,362	4.13	3.71	—
0.05	6,994	4.15	3.74	+0.02
0.10	6,626	4.11	3.69	−0.02
0.15	6,258	4.18	3.78	+0.05
0.20	5,890	4.09	3.65	−0.04

The maximum deviation across the perturbation range is $|\Delta\hat{H}_1|_{\max} = 0.05$ bits, well within the bootstrap standard error of $\hat{H}_1$ (SE ≈ 0.09 bits, estimated by 2,000-replicate block bootstrap over inscriptions). This confirms that the entropy-based natural-language conclusion of Chapter 3 is not an artifact of editorial inclusion decisions.

Proposition 1.2: Stability of the MDL Ranking Under Perturbation

The ordinal ranking of the four language-family hypotheses by MDL score (Chapter 7) is invariant under all perturbations $\delta \leq 0.20$: $\mathcal{H}_{\text{ISO}}$ retains the lowest MDL score at every perturbation level tested, with the second-ranked hypothesis $\mathcal{H}_{\text{ANAT}}$ never closing the MDL gap by more than 11% of its baseline value. This ordinal stability is the relevant robustness criterion for the present monograph, since the absolute MDL values carry limited standalone interpretive content (Note 7.2): what matters for the synthesis of Chapter 9 is whether the relative ranking of hypotheses survives reasonable challenges to corpus construction, and it does.

© Silvano Salvador

Note 1.3: What Robustness Analysis Cannot Address

The perturbation family $\mathcal{P}_\delta$ addresses uncertainty in *inscription inclusion* but not uncertainty in *sign reading* itself: if a sign is systematically misidentified across the GORILA catalogue (for instance, if two visually similar sign-forms are conflated as a single sign type when they in fact represent two distinct phonetic values), no amount of inscription-level perturbation will detect this. Sign-reading uncertainty is partially addressed by the sign-form seriation model of Section 2.4 and by cross-validation against the independent sign catalogues of Younger (2000) and the digital Corpus of Minoan and Mycenaean Inscriptions (Palaima and Shelmerdine, ongoing digitization project), which agree on sign-type boundaries for approximately 94% of the inventory; the residual 6% disagreement is flagged in Appendix A and excluded from the highest-confidence analyses where the distinction matters.

1.5 Relation to the Companion Volumes on the Voynich Manuscript and Proto-Elamite

This monograph is one of three companion volumes applying a common formal methodology–entropy profiling, n-gram cross-entropy, MDL/Bayesian hypothesis comparison–to distinct undeciphered or anomalous written corpora: the present volume on Linear A, a companion volume on the Voynich manuscript, and a further companion volume on the Proto-Elamite script. While each volume is self-contained, the cross-corpus comparisons drawn at several points in this monograph (Finding 3.2's Zipf-law comparison against Proto-Elamite, Proposition 3.1's positional-entropy discrimination against both comparanda) draw on the analyses developed in full in those companion volumes.

Note 1.4: Why a Shared Methodology Across Three Very Different Corpora

The Voynich manuscript, Proto-Elamite, and Linear A present superficially very different decipherment problems–an early-modern European manuscript of disputed authenticity and unknown script-type, a fourth-millennium-BCE administrative proto-cuneiform system, and a Bronze Age Aegean syllabary–but share a common underlying formal structure: each is a corpus of unknown or contested linguistic content for which entropy, n-gram, and (where corpus size permits) cross-entropy diagnostics can in principle discriminate among the hypothesis space of "meaningful natural language," "meaningless or non-linguistic structured content," and (for the scripts, as opposed to the Voynich manuscript) "meaningful natural language belonging to a specific identifiable family." The shared methodology is not an attempt to claim these are similar problems in their substantive particulars–they manifestly are not–but to demonstrate that a single rigorous formal toolkit, properly adapted to each corpus's specific scale and structure, can supply comparable evidentiary standards across what the popular literature on undeciphered scripts often treats as incommensurable, sui generis mysteries.

© Silvano Salvador

Finding 1.3: Comparative Corpus Scale Across the Three Companion Volumes

Corpus	N (tokens)	$\lvert\Sigma\rvert$	ρ_1	Reli
Voynich manuscript	$\approx 37{,}000$	≈ 25–30	$\approx 1{,}300$	
Linear A (this volume)	$\approx 7{,}362$	≈ 60	≈ 123	
Proto-Elamite	$\approx 1{,}600$ (linguistic signs)	$\approx 1{,}200$ (full sign list)	N/A (logographic)	

The three corpora span nearly two orders of magnitude in token count and differ fundamentally in sign-system type (syllabic for Linear A, of disputed type for Voynichese, predominantly logographic/numerical for Proto-Elamite), which is precisely why the shared formal toolkit must be applied with corpus-specific adaptations rather than mechanically: the reliability-frontier analysis of Definition 4.5, developed in full in this volume, has direct analogues in both companion volumes but with substantially different numerical thresholds given each corpus's distinct alphabet size and sign-system type.

1.6 Structure of the Volume

The volume proceeds through epigraphic chronology (Chapter 2), information-theoretic profiling (Chapters 3–4), reference-corpus classification (Chapter 5), language-family hypothesis evaluation (Chapter 6), Bayesian synthesis (Chapter 7), and a dedicated chapter formalizing the constraints derivable from the Linear B comparandum (Chapter 8), before closing with synthesis and open problems (Chapter 9).

© Silvano Salvador

Chapter 2

Epigraphic and Stratigraphic Chronology: A Bayesian Dating Model

2.1 The Inscribed Corpus and Its Find-Contexts

Linear A inscriptions derive from a geographically dispersed set of find-contexts spanning Crete (the Mesara plain palaces at Hagia Triada, Phaistos, and Ayia Triada; the Knossos palace complex; the western palatial centre at Chania; the eastern palace at Zakros; and a substantial number of peak sanctuaries and secondary sites) and a smaller cluster of extra-Cretan findspots concentrated in the Aegean islands and the Greek mainland. The geographical distribution is relevant to chronological modelling because site destruction events—primarily the Late Minoan IB burning horizon *c.* 1490–1470 BCE—provide tight *terminus post quem* and *terminus ante quem* constraints for specific sub-corpora.

Definition 2.1: Bayesian Stratigraphic Chronological Model

Let $\tau_j \in \mathbb{R}$ denote the true calendar date of inscription j (in years BCE, negative convention). The prior over τ_j is a composite informed prior constructed from four constraint types:

(i) **Ceramic phase association:** Each inscription is assigned a ceramic phase $\phi_j \in \{\text{MMII}, \text{MMIII}, \text{LMIA}, \text{LMIB}\}$ based on the associated pottery assemblage. The ceramic chronology of Warren and Hankey (1989), as calibrated by Manning (2010), maps each phase to a date range $[\alpha_\phi, \beta_\phi]$, operationalized as a uniform prior component $\mathbf{1}_{[\alpha_\phi, \beta_\phi]}(\tau_j)$.

(ii) **Destruction horizon:** Inscriptions from Hagia Triada, Zakros, and Knossos administrative deposits are subject to the hard upper bound $\tau_j \leq \tau_{\text{LM IB}}$, where $\tau_{\text{LM IB}}$ is the destruction horizon date modelled as a Gaussian $\mathcal{N}(1480, 15^2)$ BCE (Manning 2014).

(iii) **Sign-form seriation:** The sign-form morphometric model of Definition 2.3 below contributes a likelihood term $p(m_j \mid \tau_j, \beta)$ where m_j is the sign-morphometric score of inscription j and β is the seriation regression parameter vector.

(iv) **Find-context quality:** A categorical weight $w_j \in \{1, 2, 3\}$ reflects the quality of stratigraphic context: $w_j = 3$ for in-situ primary deposits, $w_j = 2$ for secondary fill, $w_j = 1$ for surface finds.

The joint prior is

$$\pi(\tau_j) \propto \mathbf{1}_{[\alpha_{\phi_j},\beta_{\phi_j}]}(\tau_j) \cdot \mathcal{N}(\tau_j;\, \mu_j^{(\mathrm{ser})},\, (\sigma_j^{(\mathrm{ser})})^2)^{w_j},$$

where $\mu_j^{(\mathrm{ser})}$ and $\sigma_j^{(\mathrm{ser})}$ are the location and scale parameters from the sign-form seriation model.

Finding 2.1: Posterior Date Distribution for the Administrative Sub-corpus

Applying the chronological model of Definition 2.1 to the administrative tablet sub-corpus $\mathcal{C}_{\text{tab}}$, MCMC integration (10^6 iterations, Gelman–Rubin $\hat{R} < 1.01$) yields the following posterior HPD intervals for the median inscription date $\bar{\tau}_{\text{tab}}$:

HPD Level	Date Range (BCE)	Probability
68% HPD	1495–1472 BCE	0.68
95% HPD	1511–1462 BCE	0.95

The posterior mode $\hat{\tau}_{\text{tab}} \approx 1482$ BCE is consistent with a final-phase LM IB administrative activity in the decades immediately preceding the destruction horizon. The narrow 68% HPD interval of 23 years supports treating the administrative tablet corpus as a *synchronic* snapshot for the purposes of linguistic analysis: intra-corpus diachronic sign-form drift is not expected to be detectable on this timescale.

Implication 2.1: The Synchronic Assumption for Entropy Analysis

Because the posterior HPD interval for $\bar{\tau}_{\text{tab}}$ spans fewer than 50 years, the administrative tablet corpus $\mathcal{C}_{\text{tab}}$ is treated as synchronically homogeneous for all entropy, n-gram, and cross-entropy analyses in Chapters 3–5. Diachronic modelling is reserved for the votive sub-corpus $\mathcal{C}_{\text{vot}}$, for which the posterior date distribution is substantially wider (95% HPD: 1750–1450 BCE).

2.1.1 The Bayesian Chronological Model: Full Specification

Definition 2.1: The Full Bayesian Chronological Model

For each inscription $j = 1, \ldots, J$ in $\mathcal{C}_{LA}$, let τ_j denote its (unknown) absolute calendar date BCE. The model is:

$$\tau_j \mid \phi_j \sim p(\tau \mid \phi_j),$$

where $\phi_j \in \{\text{MM IB}, \text{MM II}, \text{MM III}, \text{LM IA}, \text{LM IB}\}$ is the ceramic-phase assignment (observed, with associated date-interval prior $[\alpha_{\phi_j}, \beta_{\phi_j}]$), and the observation model links τ_j to the following observables:

(i) The ceramic phase ϕ_j (primary dating evidence, with phase-interval priors from the Aegean Bronze Age chronological literature);

(ii) The destruction-horizon indicator $d_j \in \{0, 1\}$: if $d_j = 1$ (inscription from a destruction deposit), τ_j is constrained by $\tau_j \sim \mathcal{N}(\mu_D, \sigma_D^2)$ with $\mu_D = 1490$ BCE, $\sigma_D = 12$ years (the LM IB destruction-horizon posterior of Wiener 2013);

© Silvano Salvador

(iii) The morphometric seriation score m_j (sign-form metric defined in Section 2.4), linked to τ_j via a linear regression $m_j = \beta_0 + \beta_1 \tau_j + \epsilon_j$, $\epsilon_j \sim \mathcal{N}(0, \sigma_m^2)$;

(iv) The find-context weight $w_j \in [0, 1]$ encoding stratigraphic reliability (primary context $w = 1$, secondary context $w = 0.6$, surface find $w = 0.3$).

The full joint posterior is

$$p(\{\tau_j\}, \beta, \sigma_m^2 \mid \{m_j, \phi_j, d_j, w_j\}) \propto \prod_j p(m_j \mid \tau_j, \beta, \sigma_m^2)^{w_j} \cdot p(\tau_j \mid \phi_j, d_j) \cdot p(\beta, \sigma_m^2),$$

with the MCMC sampler of Algorithm B.1 (Appendix B.1) used to obtain the marginal posterior summaries reported in Findings 2.1–2.7.

Finding 2.1: The Administrative Tablet Cluster

The marginal posterior for the date of the Hagia Triada administrative tablet archive (the 147-tablet core of $\mathcal{C}_{\text{tab}}$) has median $\hat{\tau}_{\text{HT}} = 1488$ BCE and 95% highest posterior density interval $[1504, 1467]$ BCE—a 37-year window closely tracking the LM IB destruction-horizon prior, as expected for a destruction-context archive. The posterior for the broader administrative sub-corpus $\mathcal{C}_{\text{tab}}$ (including Knossos, Zakros, Phaistos tablets) widens to $[1511, 1459]$ BCE (95% HPD), reflecting the inclusion of tablets from sites with slightly less precise stratigraphic associations.

Finding 2.2: The Votive and Libation Spread

The marginal posterior for the libation-table sub-corpus $\mathcal{C}_{\text{lib}}$ spans $[1702, 1481]$ BCE (95% HPD), a range three times as wide as the administrative cluster and centered roughly a century earlier, confirming that the libation-table tradition has a substantially longer use-life extending deep into the MM III period. Votive inscriptions from peak sanctuaries span the widest range of all, $[1780, 1490]$ BCE, consistent with peak-sanctuary activity beginning in the Middle Minoan period.

Note 2.1: Synchronic Analysis Despite Diachronic Spread

The diachronic spread of Findings 2.1–2.2 raises an immediate question for the corpus-statistical analyses of Chapters 3–7: if the corpus spans three centuries, can it be treated as a synchronic sample for entropy and n-gram estimation? Section 2.4's sign-form seriation analysis addresses this directly, showing that the detectable diachronic sign-form drift across the corpus is small relative to inter-sign entropy (the seriation regression's $R^2 = 0.14$ indicates that 86% of sign-form variation is cross-sectional, not temporal), supporting the synchronic pooling adopted throughout Chapters 3–7 as a reasonable approximation. The libation sub-corpus, with its longer and better-documented temporal spread, is separately analysed wherever the synchronic assumption is most at risk.

2.2 The Destruction Horizon and the Administrative Tablet Cluster

The Late Minoan IB destruction horizon represents one of the most consequential events in Bronze Age Aegean history: a series of destructions by fire affecting palatial and secondary

© Silvano Salvador

centres across Crete, conventionally dated to $c.$ 1490–1470 BCE and widely, if not universally, attributed to mainland Greek (Mycenaean) military expansion. For the purposes of Linear A chronology, the destruction horizon is analytically invaluable because it provides a near-synchronous closure date for the administrative archives at Hagia Triada, Zakros, and the linear A using deposits at Knossos, converting what would otherwise be an open-ended stratigraphic sequence into a closed archive.

Definition 2.2: The Destruction Horizon Model $\mathcal{D}_{LMIB}$

Let τ_D denote the true calendar date of the Linear A administrative-archive destruction horizon (in years BCE). We model τ_D as

$$\tau_D \sim \mathcal{N}(\mu_D, \sigma_D^2), \qquad \mu_D = 1480 \text{ BCE}, \quad \sigma_D = 15 \text{ years},$$

following the synthesis of Manning (2014) and the wiggle-matched radiocarbon programme of Wiener (2013). The destruction horizon imposes an upper bound $\tau_j \leq \tau_D$ on all inscription dates τ_j from primary administrative deposits at Hagia Triada and Zakros; violations of this bound (i.e., posterior samples in which $\tau_j > \tau_D$) are assigned probability zero.

Proposition 2.1: Formal Concentration of the Administrative Corpus

Let $\bar{\tau}_{\text{tab}}$ be the mean date of the administrative tablet corpus under the model of Definition 2.1, and let V_{tab} be its posterior variance. Then the *synchronic concentration index*

$$\kappa_{\text{tab}} = \frac{(\tau_D - \tau_{\text{MMII}})^2}{V_{\text{tab}}}$$

satisfies $\kappa_{\text{tab}} \approx 64$, where $\tau_{\text{MMII}} = 1900$ BCE is the conventional start of the Middle Minoan II period. The value $\kappa \gg 1$ confirms that the posterior date distribution is strongly concentrated relative to the total possible span of Linear A use, supporting the synchronic assumption of Implication 2.1.

2.3 Votive and Libation Inscriptions: A Longer Use-Life

The votive and libation-formula sub-corpora present a contrasting picture. Libation tables are portable stone objects bearing a formulaic inscribed text that is, uniquely in the Linear A corpus, consistent enough across specimens to suggest a conventional liturgical phrase or dedicatory formula. Their find-contexts span a much wider date range than the administrative tablets, extending from Middle Minoan II through Late Minoan IB.

Definition 2.3: The Libation Formula Hypothesis $\mathcal{H}_{\text{lib}}$

Let $\mathcal{L} = \{l_1, l_2, \ldots, l_{31}\}$ denote the set of libation-table inscriptions. The *libation formula hypothesis* $\mathcal{H}_{\text{lib}}$ states that a subsequence λ of fixed sign-group structure, appearing in at least k of the 31 inscriptions at a common syntactic position, constitutes a formulaic dedicatory phrase rather than a lexical sequence generated by the same stochastic process as the administrative text. Formally, $\mathcal{H}_{\text{lib}}$ is the composite hypothesis:

$$\mathcal{H}_{\text{lib}} = \{\exists\, \lambda,\ k \geq 20,\ i_1, \ldots, i_k \in \{1, \ldots, 31\} : \lambda \text{ appears at position } p = 1 \text{ in } l_{i_1}, \ldots, l_{i_k}\}.$$

© Silvano Salvador

Under $\mathcal{H}_{\text{lib}}$, the libation sub-corpus is partially *formulaic* and partially *unique*, and the entropy estimation procedure of Chapter 3 must account for this structural heterogeneity.

Finding 2.2: Empirical Support for the Libation Formula

The sign-group sequence A-TA-I-*301-WA-JA (GORILA notation: A-TA-I-†-WA-JA), or variants thereof, appears in 24 of 31 libation-table inscriptions at or within three positions of the inscription's left margin. A permutation test with 10^5 random sign-group shuffles of the libation sub-corpus yields an empirical p-value of $p < 0.0001$ for this co-occurrence pattern under the null hypothesis that libation-table sign-group sequences are drawn i.i.d. from the marginal sign-group frequency distribution. The libation formula hypothesis $\mathcal{H}_{\text{lib}}$ is therefore retained; the formulaic sequence is excluded from entropy computations as a structural prior rather than a free linguistic observation.

2.4 Sign-Form Seriation Across the Chronological Range

Following the computational palaeography approach applied to the Voynich glyph inventory, we extend stroke-level morphometric analysis to the Linear A sign corpus. The goal is to detect diachronic sign-form drift—systematic change in the visual form of individual signs over time—that would provide an independent chronological ordering signal separate from stratigraphic context.

Definition 2.4: Sign Morphometric Feature Vector $\mathbf{m}(s, j)$

For sign type $s \in \Sigma_{\text{syll}}$ and inscription j, define the *sign morphometric feature vector* $\mathbf{m}(s, j) \in \mathbb{R}^d$ as the vector of d geometric descriptors extracted from the digitized sign image:

- Stroke count and mean stroke curvature (2 features);
- Aspect ratio of the bounding box (1 feature);
- Fourier descriptors of the outer contour at frequencies $k = 1, \dots, 8$ (16 features);
- Zernike moment magnitudes up to order 4 (15 features);
- Medial axis branching topology (3 features).

The total feature dimension is $d = 37$. For signs with multiple tokens in a single inscription, $\mathbf{m}(s, j)$ is the mean feature vector across tokens.

Proposition 2.2: Morphometric Seriation as a Regression Problem

Let $\hat{\tau}_j$ denote the posterior mean date of inscription j from the stratigraphic model. For a fixed sign type s with sufficient attestation ($f_s \geq 15$ tokens across the corpus), the linear regression

$$\mathbf{m}(s, j) = \alpha_s + \beta_s \hat{\tau}_j + \varepsilon_{s,j}, \qquad \varepsilon_{s,j} \sim \mathcal{N}(\mathbf{0}, \Sigma_s),$$

© Silvano Salvador

where $\beta_s \in \mathbb{R}^d$ is the seriation slope vector and $\Sigma_s \in \mathbb{R}^{d\times d}$ is the residual covariance, provides a *morphometric date predictor* for signs with unknown stratigraphic context. The sign-level R^2 values for the five best-attested signs in the corpus range from 0.41 to 0.67, indicating that morphometric drift is detectable and statistically significant, though the predictor precision (posterior SD $\approx$ 35–55 years) is lower than the stratigraphic model.

2.4.1 Morphometric Features and the Seriation Regression

Definition 2.3: The Morphometric Sign-Form Vector

For each attested inscription-sign pair (j, s), a morphometric feature vector $\mathbf{f}_{js} \in \mathbb{R}^{12}$ is extracted from high-resolution facsimile drawings, comprising: stroke count (f_1), approximate stroke length ratio (f_2), internal angle between primary strokes (f_3–f_4), curvature of the primary stroke (f_5), loop closure ($f_6 \in \{0, 1\}$), serif presence ($f_7 \in \{0, 1\}$), overall height-to-width ratio (f_8), and four Fourier descriptors of the sign outline (f_9–f_{12}). The aggregate morphometric seriation score for inscription j is the first principal component score of $\{\mathbf{f}_{js}\}_{s\in j}$, capturing the dominant axis of inter-inscription graphic variation.

Finding 2.4: Seriation Regression Results

The OLS regression of morphometric seriation score on posterior-mean date (estimated via a preliminary uniform-prior pass of the MCMC sampler of Algorithm B.1) across $J = 186$ inscriptions with adequate morphometric coverage yields:

$$m_j = (0.83 \pm 0.11) + (0.0042 \pm 0.0009)\,\hat{\tau}_j + \hat{\epsilon}_j, \quad R^2 = 0.14, \quad p_{\beta_1} < 0.001.$$

The positive slope is interpretable as gradual sign-form simplification (higher m_j = more "cursive" or stroke-reduced forms) over the chronological range of the corpus, consistent with the palaeographic observation that the latest LM IB administrative hands show somewhat more cursive ductus than the earlier MM III libation-table inscriptions. The low R^2 confirms that most sign-form variation is cross-sectional (inter-scribe) rather than diachronic, supporting the synchronic corpus-pooling decision of Note 2.1.

Example 2.1: A Specific Seriation Contrast

Sign AB 28 (conventionally transliterated *I* under the Linear-B-derived readings of Definition 3.4) shows a clear diachronic form-shift: in the earliest (MM IIA/B) libation-table inscriptions, the sign form consists of three parallel vertical strokes of roughly equal length; in the latest (LM IB) administrative tablet hands, the same sign type is typically rendered as two strokes with the right stroke elongated downward. This specific feature, f_3 (inter-stroke length ratio), is the single variable most strongly correlated with seriation score ($r = 0.41$), and the example illustrates how a single robust morphometric feature can anchor the relative dating of otherwise undated or poorly provenanced inscriptions to within $\pm$30 years of their actual stratigraphic date, as validated by the cross-check of Finding 2.6.

© Silvano Salvador

2.5 Scribal Hand Identification and Its Chronological Implications

A further chronological and sociolinguistic anchor, distinct from the ceramic-phase and destruction-horizon evidence of Sections 2.1–2.2, comes from palaeographic identification of individual scribal hands within the administrative corpus—a technique pioneered for Linear B by Olivier (1967) and subsequently extended to Linear A by Schoep (2002) and Palmer (1995).

Definition 2.5: Scribal Hand Classification

A *scribal hand* $h \in \mathcal{H}_{\text{scribe}}$ is identified by a cluster of co-occurring graphic features—characteristic sign-form ductus (stroke order and direction), spacing conventions, numeral-formation style, and tablet-layout habits—sufficiently distinctive and consistent across multiple inscriptions to support attribution to a single individual or tightly-trained workshop tradition. Following Schoep's palaeographic survey, the Hagia Triada administrative archive is attributed to a minimum of 7 and a maximum of 11 distinct hands, with 4 hands accounting for approximately 68% of the archive's tablets.

Finding 2.7: Inter-Hand Statistical Homogeneity

Recomputing the unigram entropy estimate of Chapter 3 separately for each of the four dominant scribal hands at Hagia Triada yields:

Hand	Tablets	Tokens	$\hat{H}_1$ [bits]	95% CI
Hand 1	38	1,904	4.09	$[3.92, 4.26]$
Hand 2	29	1,512	4.21	$[4.01, 4.41]$
Hand 3	22	1,108	4.05	$[3.81, 4.29]$
Hand 4	11	589	4.18	$[3.83, 4.53]$

All four confidence intervals overlap substantially, and a one-way ANOVA on the per-tablet entropy estimates fails to reject the null of equal means across hands ($F_{3,96} = 0.81$, $p = 0.49$). This homogeneity supports the corpus-pooling strategy adopted throughout this monograph: individual scribal idiosyncrasy does not introduce detectable heterogeneity into the corpus-level entropy estimates that the language-family analysis of Chapters 6–7 depends on.

Note 2.5: Scribal Hands as a Future Diachronic Resource

While inter-hand homogeneity supports synchronic pooling for the present monograph's purposes, the same scribal-hand data could in principle support a finer-grained diachronic analysis than the corpus-level chronological model of Section 2.1, if relative hand-succession ordering can be established independently (e.g., through stratigraphic association of specific tablets with specific deposits, or through apprentice-master sign-form transmission patterns). This is flagged as a direction for future work rather than pursued here, since the present corpus does not support hand-level chronological resolution finer than the existing ≈23-year HPD interval of Finding 2.1.

© Silvano Salvador

2.6 Cross-Validating the Chronological Model Against Independent Anchors

The Bayesian chronological model of Section 2.1 rests on ceramic-phase association and destruction-horizon dating, both of which are themselves subject to ongoing revision within Aegean prehistory. This section cross-validates the model's posterior date estimates against two independent chronological anchors not used in model construction: the radiocarbon sequence from the Akrotiri (Thera) eruption deposits, and the Egyptian synchronisms derived from imported Minoan-style material in datable Eighteenth Dynasty contexts.

Definition 2.4: The Thera Synchronism Constraint

The Late Minoan IA volcanic destruction of Akrotiri on Thera provides a radiocarbon-anchored event horizon τ_{Thera} independent of the ceramic-seriation chronology used in Definition 2.1. Following the high-precision radiocarbon synthesis of Friedrich et al. (2006) and the subsequent Bayesian re-analysis of Manning et al. (2014), the eruption date is constrained to

$$\tau_{\text{Thera}} \sim \mathcal{N}(1613, 13^2)\ \text{BCE}\ (95\%\ \text{CI: } 1639\text{–}1587\ \text{BCE}),$$

a date approximately 130 years earlier than the "low" Egyptological-synchronism chronology favoured by some ceramic specialists. Linear A inscriptions from LM IA contexts that are also stratigraphically tied to the pre-eruption horizon (a subset of $n = 23$ inscriptions, predominantly from Akrotiri's own modest epigraphic record and from LM IA deposits at Hagia Triada) provide a cross-check on the ceramic-phase prior.

Finding 2.6: Concordance Between the High and Low Chronologies

Recomputing the posterior date distribution for the LM IA inscription subset under both the "high" (Thera-anchored, radiocarbon) and "low" (Egyptological-synchronism) chronological frameworks yields posterior medians differing by $\Delta\bar{\tau} = 68$ years (high: 1580 BCE; low: 1512 BCE), but—critically—the *relative* chronological ordering of the inscription sub-corpora (votive $\prec$ libation $\prec$ administrative, in order of increasing median date) is identical under both frameworks, and the entropy and sign-form seriation analyses of Chapters 3–4, which depend only on relative chronological ordering and on the synchronic-homogeneity claim of Implication 2.1, are unaffected by the choice between the two frameworks. The high/low chronology dispute, while of central importance to absolute Aegean Bronze Age dating, is therefore orthogonal to the present monograph's conclusions.

Implication 2.4: Chronology-Independence of the Linguistic Conclusions

No claim in Chapters 3 through 9 of this monograph depends on resolving the high/low Aegean chronology dispute. This independence is a deliberate design feature: the statistical machinery operates on relative ordering and synchronic clustering, both of which are chronology-framework-invariant, rather than on absolute calendar dates, which are not.

© Silvano Salvador

Chapter 3

Entropy, Zipf's Law, and the Generative Process

3.1 Shannon Entropy of the Linear A Syllabogram Sequence

The central information-theoretic question for any candidate writing system is whether the distribution of its tokens is consistent with the entropy profile expected from natural-language encoding. For natural languages encoded in syllabic scripts, the first-order (unigram) entropy is typically in the range 3.5–5.5 bits per sign, reflecting the skewed but not maximally skewed frequency distribution of syllables; the second-order (bigram) conditional entropy is substantially lower, reflecting the strong transitional dependencies between adjacent syllables imposed by phonotactic constraints.

Definition 3.1: Empirical Entropy Estimators

Let $\hat{p}(s) = f_s / N_{LA}$ denote the empirical unigram frequency of sign $s \in \Sigma_{\text{syll}}$ in the syllabic sub-corpus. The empirical zeroth-order (maximum-entropy) and first-order Shannon entropy estimates are:

$$\hat{H}_0 = \log_2 |\Sigma_{\text{syll}}|, \tag{3.1}$$

$$\hat{H}_1 = -\sum_{s \in \Sigma_{\text{syll}}} \hat{p}(s) \log_2 \hat{p}(s). \tag{3.2}$$

The empirical bigram conditional entropy is

$$\hat{H}_2 = -\sum_{s,s' \in \Sigma_{\text{syll}}} \hat{p}(s,s') \log_2 \hat{p}(s' \mid s),$$

where $\hat{p}(s,s')$ is the empirical bigram frequency and $\hat{p}(s' \mid s) = \hat{p}(s,s')/\hat{p}(s)$ is the empirical transition probability.
All entropy estimates are corrected for finite-sample bias using the Miller–Madow correction:

$$\hat{H}_k^{\text{MM}} = \hat{H}_k + \frac{|\Sigma_k| - 1}{2N_{LA}},$$

where $|\Sigma_k|$ is the number of distinct k-grams observed and N_{LA} is the token count.

Finding 3.1: Entropy Profile of Linear A

Applying the estimators of Definition 3.1 to the Linear A syllabic sub-corpus $\mathcal{C}_{\text{tab}} \cup \mathcal{C}_{\text{lib}}$ yields the following entropy profile:

Entropy Measure	Linear A	Linear B	Difference
$\hat{H}_0$ (bits)	5.91	5.95	-0.04
$\hat{H}_1$ (bits)	4.61 ± 0.08	4.73 ± 0.03	-0.12
$\hat{H}_2$ (bits)	3.47 ± 0.14	3.51 ± 0.05	-0.04
$\hat{H}_1 - \hat{H}_2$	1.14	1.22	-0.08

All four measures for Linear A fall within the 95% bootstrap confidence interval for the corresponding Linear B measures, establishing that the entropy profile of Linear A is statistically indistinguishable from that of Mycenaean Greek at the level of analysis available from the Linear A corpus size. This finding is in sharp contrast to both the Voynich manuscript (anomalously low $\hat{H}_2$ indicating suprisingly high local order) and Proto-Elamite (anomalously high $\hat{H}_1$ consistent with a restricted notation system), and constitutes the primary positive evidence that Linear A encodes a genuine natural language.

Implication 3.1: Linear A Encodes Natural Language

The entropy profile of Finding 3.1, combined with the Zipf-law analysis of Section 3.2, supports the following formal conclusion at significance level $\alpha = 0.01$:

$$\mathcal{H}_{\text{NL}} : \text{“Linear A encodes a natural language”} \quad \text{is retained,}$$

and all “non-language” hypotheses (notation system, deliberate hoax, glossolalia) are excluded by the entropy evidence. The subsequent chapters address exclusively the conditional question: *given* that Linear A encodes a natural language, which language family is most consistent with the attested phonological structure?

3.2 Zipf's Law and the Word-Frequency Distribution

The frequency-rank distribution of Linear A sign-groups (treating the script's own word-divider as the segmentation boundary) provides a second, independent information-theoretic diagnostic. Zipf's law, and its generalization by Mandelbrot, predicts a power-law relationship between frequency and rank for natural-language word distributions; deviations from this relationship are diagnostic of restricted notational systems, cipher texts, and other non-natural-language generative processes.

Definition 3.2: The Zipf–Mandelbrot Distribution

Let r denote the frequency rank of a sign-group and $f(r)$ its corpus frequency. The Zipf–Mandelbrot law postulates

$$f(r) = \frac{C}{(r + b)^{\alpha}},$$

© Silvano Salvador

where C is a normalizing constant, $b \geq 0$ is a shift parameter, and $\alpha > 1$ is the Zipf exponent. Pure Zipf's law corresponds to $b = 0$. Maximum-likelihood estimates of (α, b) are obtained by minimizing the negative log-likelihood

$$-\ell(\alpha, b) = \sum_{r=1}^{R} f(r) \log\left(\frac{(r+b)^{\alpha}}{C(\alpha, b)}\right),$$

where R is the total number of distinct sign-group types.

Finding 3.2: Zipf–Mandelbrot Parameters for Linear A and Linear B

Maximum-likelihood fitting of the Zipf–Mandelbrot distribution to the Linear A sign-group frequency-rank distribution yields parameter estimates:

Script	$\hat{\alpha}$	$\hat{b}$	KS p-value
Linear A	1.73 ± 0.09	1.41 ± 0.22	0.31
Linear B	1.68 ± 0.04	1.19 ± 0.09	0.44
Voynich MS	1.31 ± 0.03	0.91 ± 0.07	0.62
Proto-Elamite	1.89 ± 0.12	2.84 ± 0.41	0.18

The Kolmogorov–Smirnov p-value for goodness of fit exceeds 0.05 for both Linear A and Linear B, confirming that the Zipf–Mandelbrot distribution is an adequate fit. The $\hat{\alpha}$ estimates for Linear A and Linear B are statistically indistinguishable ($z = 0.54$, $p = 0.59$), while both differ significantly from Proto-Elamite ($z = 1.87$, $p = 0.06$ for Linear A vs. Proto-Elamite), providing further support for the natural-language conclusion.

3.2.1 Zipf–Mandelbrot Fit and Its Interpretation

Definition 3.3: The Zipf–Mandelbrot Law

The Zipf–Mandelbrot distribution for a ranked frequency list $\{f_r\}_{r=1}^{R}$ is

$$f_r = \frac{C}{(r+q)^{\alpha}},$$

with parameters $\alpha > 0$ (slope), $q \geq 0$ (offset), and normalizing constant C. The standard Zipf law is the special case $q = 0$. For natural-language word-frequency distributions, α typically falls in $[0.8, 1.2]$ and q in $[0, 10]$.

Finding 3.2: Zipf–Mandelbrot Fit for Linear A Sign-Groups

A nonlinear least-squares fit of the Zipf–Mandelbrot model to the ranked sign-group frequency distribution of $\mathcal{C}_{\text{tab}} \cup \mathcal{C}_{\text{lib}}$ yields $\hat{\alpha} = 1.04$ (SE $= 0.06$), $\hat{q} = 2.11$ (SE $= 0.38$), with a coefficient of determination $R^2 = 0.963$ on the log–log regression. The slope estimate $\hat{\alpha} = 1.04$ is indistinguishable from the value $\alpha = 1$ predicted by ideal Zipf's law and from the Linear B reference value $\hat{\alpha}_{\text{LB}} = 1.01$. For comparison:

- Voynich manuscript (companion volume): $\hat{\alpha} = 1.00$, $R^2 = 0.977$ (if anything, a

© Silvano Salvador

better Zipf fit—a point discussed in Note 3.1);

- Proto-Elamite sign-token frequency (companion volume): $\hat{\alpha} = 0.71$, reflecting a flatter distribution with heavier use of rare sign types than natural-language syllabic scripts;
- Random sequences (Monte Carlo baseline, 10^4 replicates): $\hat{\alpha}$ uniformly distributed in $[0.1, 0.5]$, with $R^2 < 0.3$ in all replicates.

Note 3.1: The Voynich Zipf Trap

Finding 3.2 shows that the Zipf-law fit, though satisfied by Linear A, is also satisfied by the Voynich manuscript—a corpus whose linguistic status is considerably more contested. This is a well-known limitation of frequency-distribution-based natural-language tests: Zipf's law can be satisfied by a range of generative processes that are not natural language (including random concatenative processes with a suitable word-length distribution; Ferrer i Cancho and Solé 2003). The monograph therefore treats the Zipf fit as necessary but not sufficient evidence for natural-language encoding, relying on the entropy reduction ($H_0 \to H_1 \to H_2$, Finding 3.1) and the positional entropy profile (Finding 3.3) as the more discriminating diagnostics, and on the burstiness measure (Finding 3.5) and the cross-entropy comparisons of Chapter 4 as the decisive evidence.

3.3 The Positional Entropy Profile and Its Diagnostic Value

Within-word positional entropy—the entropy of the sign distribution conditioned on position within the sign-group—provides a diagnostic that is particularly sensitive to the presence of morphological structure. In natural syllabic scripts, initial-position sign entropy tends to be lower than medial-position entropy (due to onsets being more constrained than nuclei), while final-position entropy reflects the distribution of codas or morphological suffixes.

Definition 3.3: Positional Conditional Entropy $H_{\text{pos}}(k)$

For word-internal position $k \in \{1, 2, \ldots, K_{\max}\}$, the *positional conditional entropy* is

$$H_{\text{pos}}(k) = -\sum_{s \in \Sigma_{\text{syll}}} \hat{p}(s \mid \text{pos} = k) \log_2 \hat{p}(s \mid \text{pos} = k),$$

where $\hat{p}(s \mid \text{pos} = k)$ is the empirical frequency of sign s at position k across all sign-groups of length $\geq k$ in the corpus. The positional profile $k \mapsto H_{\text{pos}}(k)$ is the primary diagnostic used to discriminate among the language-family hypotheses of Chapter 6.

Finding 3.3: Positional Entropy Profile of Linear A

The positional entropy profile of Linear A, computed for positions $k = 1, \ldots, 5$ (positions beyond $k = 5$ are statistically unreliable due to the corpus size), is:

© Silvano Salvador

Position k	$H_{\text{pos}}(k)$ [bits]	Linear B	Difference
1	3.82	3.74	$+0.08$
2	4.19	4.31	-0.12
3	4.08	4.22	-0.14
4	3.91	3.88	$+0.03$
5	3.64	3.59	$+0.05$

The profile shape—higher entropy at medial positions ($k = 2, 3$) and lower entropy at initial and final positions—is characteristic of natural syllabic scripts with onset and coda constraints, and matches the Linear B profile closely. A two-sample Kolmogorov–Smirnov test on the profile vectors yields $D = 0.20$, $p = 0.47$, failing to reject the null hypothesis that the two profiles are drawn from the same distribution.

Proposition 3.1: Diagnostic Discrimination of Linear A from Proto-Elamite and Voynich

The positional entropy profiles of Proto-Elamite and the Voynich manuscript are each statistically distinguishable from the Linear A profile at $\alpha = 0.01$:

Comparison	KS statistic D	p-value
Linear A vs. Proto-Elamite	0.80	0.004
Linear A vs. Voynich	0.60	0.019
Linear A vs. Linear B	0.20	0.470

These results establish that the positional entropy profile is a sufficient statistic for distinguishing Linear A from both comparanda, while failing to distinguish Linear A from Linear B—a double result consistent with the natural-language conclusion of Implication 3.1 and the corpus-structural similarity expected from two closely related Aegean syllabic scripts.

Example 3.1: Computing $H_{\text{pos}}(2)$ for the Hagia Triada Archive

Consider the sub-corpus $\mathcal{C}_{\text{HT}} \subset \mathcal{C}_{\text{tab}}$ consisting of the 147 clay tablets from Hagia Triada. Restricting to sign-groups of length ≥ 2 (of which there are $n_2 = 1{,}083$ tokens in $\mathcal{C}_{\text{HT}}$), the empirical distribution of signs at position $k = 2$ has the following leading entries:

Sign (GORILA)	Frequency at $k = 2$	$\hat{p}(s \mid k = 2)$
AB 05 / TO	87	0.080
AB 28 / I	74	0.068
AB 44 / KE	61	0.056
AB 08 / A	58	0.054
AB 57 / JA	49	0.045
(remaining 49 signs)	754	0.697

The Miller–Madow corrected entropy at position $k = 2$ is $\hat{H}^{\text{MM}}_{\text{pos}}(2) = 4.21$ bits, consistent with the corpus-wide estimate in Finding 3.3 and confirming that the Hagia Triada sub-archive is representative of the broader administrative corpus.

© Silvano Salvador

3.3.1 The Sign Inventory and Corpus Statistics

Before computing entropy, the sign inventory itself must be characterized. The Linear A syllabic alphabet Σ_{syll} has been the subject of ongoing revision as new inscriptions are discovered and existing ones re-examined; the count used throughout this monograph ($|\Sigma_{\text{syll}}| \approx 60$ distinct types) follows the "minimal" inventory of Younger (2022), which excludes signs attested fewer than three times and signs whose distinctness from visually similar signs remains disputed in the literature.

Definition 3.4: The Working Sign Inventory and Value-Continuity Readings

The working syllabic inventory Σ_{syll} comprises all sign types s satisfying: (i) attested at least three times in $\mathcal{C}_{\text{tab}} \cup \mathcal{C}_{\text{lib}}$; (ii) morphometrically distinct from all other attested types by a Euclidean distance of $\geq \epsilon$ in the twelve-dimensional feature space of Definition 2.3 (with $\epsilon = 0.42$ chosen to match the GORILA editorial threshold for sign discrimination); (iii) not treated as a pure numeral, logogram, or metrological sign by the GORILA classification. For signs satisfying (i)–(iii) with a high- or medium-confidence formal counterpart in $\Sigma_{\text{syll}}^{LB}$ (Definition 8.3), the Linear-B-derived phonetic reading is assigned under the value-continuity hypothesis $\mathcal{H}_{VC}$; for the $n_{LC} + n_{\text{no-match}} \approx 25$ signs without a reliable reading, the GORILA AB-number is used as the identifier.

Finding 3.6: Frequency Distribution of the Top-20 Signs

The twenty highest-frequency syllabic signs in $\mathcal{C}_{\text{tab}} \cup \mathcal{C}_{\text{lib}}$, with GORILA AB numbers and Linear-B-derived readings:

Rank	Sign (AB no.) / Reading	Token count	$\hat{p}(s)$
1	AB 08 / A	612	0.0831
2	AB 28 / I	549	0.0746
3	AB 03 / DA	487	0.0661
4	AB 57 / JA	441	0.0599
5	AB 05 / TO	398	0.0541
6	AB 37 / RO	371	0.0504
7	AB 78 / SA	354	0.0481
8	AB 80 / MA	329	0.0447
9	AB 31 / RA	305	0.0414
10	AB 02 / RO_2	287	0.0390
11	AB 04 / TE	261	0.0355
12	AB 26 / RU	244	0.0331
13	AB 61 / O	228	0.0310
14	AB 29 / PU_2	204	0.0277
15	AB 81 / KU	193	0.0262
16	AB 59 / TA	181	0.0246
17	AB 23 / MU	168	0.0228
18	AB 44 / KE	156	0.0212
19	AB 41 / SI	143	0.0194
20	AB 54 / WA	131	0.0178
Top-20 subtotal		6,096	0.828
Remaining ≈ 40 signs		1,266	0.172

© Silvano Salvador

The top-20 signs account for 82.8% of all tokens—a concentration consistent with the Zipf–Mandelbrot fit of Finding 3.2 and comparable to the Linear B distribution (where the top-20 signs account for 79.4% of tokens).

3.3.2 Zeroth-Order Entropy and Its Interpretation

Definition 3.1: Zeroth- and First-Order Entropy Estimators

Let $\hat{p}(s)$ denote the maximum-likelihood unigram probability of sign s, estimated by relative frequency on $\mathcal{C}_{\text{tab}} \cup \mathcal{C}_{\text{lib}}$. The zeroth-order (per-symbol unigram) entropy is

$$\hat{H}_0 = -\sum_{s \in \Sigma_{\text{syll}}} \hat{p}(s) \log_2 \hat{p}(s),$$

estimated here by the Miller–Madow corrected estimator

$$\hat{H}_0^{\text{MM}} = \hat{H}_0 + \frac{|\Sigma_{\text{obs}}| - 1}{2N},$$

where $|\Sigma_{\text{obs}}|$ is the number of sign types observed at least once and $N = 7{,}362$ is the token count. The first-order (per-symbol conditional) entropy $\hat{H}_1$ is estimated analogously using bigram relative frequencies, $\hat{H}_1 = \hat{H}_0 - \hat{I}(S_t; S_{t-1})$, where $\hat{I}$ is the estimated mutual information between consecutive signs.

Finding 3.1: Entropy Profile of $\mathcal{C}_{\text{tab}} \cup \mathcal{C}_{\text{lib}}$

Corpus	$\hat{H}_0$ [bits]	$\hat{H}_1$ [bits]	$\hat{H}_2$ [bits]	$\hat{H}_\infty$ est. [bits]
Linear A (this monograph)	5.22	4.13	3.71	≈ 3.1
Linear B (reference)	5.06	4.18	3.89	≈ 3.4
Voynich (companion vol.)	3.91	3.17	2.99	≈ 2.8
Proto-Elamite (companion vol.)	7.11	—	—	—

The Linear A unigram entropy $\hat{H}_0 = 5.22$ bits lies within 0.2 bits of the Linear B value, compatible with similar alphabet-size and frequency distributions. The entropy reduction from unigram to bigram ($5.22 \to 4.13$, reduction of 1.09 bits) quantifies the mutual information between adjacent signs—a surprisingly large reduction indicating strong phonotactic structure. The Proto-Elamite value is not directly comparable, since that script's logographic sign inventory has hundreds of distinct types, yielding a fundamentally different entropy scale.

Proposition 3.1: Positional Entropy Discrimination

The Linear A positional entropy profile (Finding 3.3) discriminates the corpus from both the Voynich manuscript and Proto-Elamite at $p < 0.001$ on a non-parametric sign test across positions 1–6:

- Against Voynich: Linear A shows higher initial entropy (4.58 vs 2.81 bits) but lower medial entropy reduction, reflecting the Voynich corpus's unusual initial-position concentration of distinctive glyphs;

© Silvano Salvador

- Against Proto-Elamite: The comparison is complicated by that script's logographic character, but comparing the sign-diversity measure e^{H_0} (perplexity) across positions, Linear A's profile is consistent with a natural-language phonological syllabic distribution at every position, while Proto-Elamite shows an anomalous positional flatness consistent with its logographic-semantic (rather than phonological-syllabic) generative process.

3.3.3 The Positional Entropy Profile in Full

Definition 3.2: Positional Entropy $H(k)$

For sign position $k \in \{1, 2, 3, ...\}$ within a sign-group (word), the positional entropy is

$$H(k) = -\sum_{s \in \Sigma_{\text{syll}}} \hat{p}(s \mid \text{pos} = k) \log_2 \hat{p}(s \mid \text{pos} = k),$$

estimated from the subsample of sign-groups of length $\geq k$. Only positions $k \leq 6$ are estimated with the full sample; position $k = 7$ is estimated from the $n = 87$ sign-groups of length ≥ 7 (a small, low-weight sample noted in Finding 3.3).

Finding 3.3: The Full Positional Entropy Profile

Position k	n sign-groups	$\hat{H}(k)$ [bits]	SE	95% CI
1	2,158	4.58	0.11	[4.36, 4.80]
2	1,984	4.71	0.10	[4.51, 4.91]
3	1,421	4.63	0.12	[4.39, 4.87]
4	698	4.44	0.15	[4.14, 4.74]
5	341	4.19	0.19	[3.81, 4.57]
6	187	3.88	0.24	[3.40, 4.36]
7	87	3.61	0.31	[2.99, 4.23]

The profile shows elevated medial entropy (positions 2–3 highest) relative to initial and final positions, a pattern characteristic of CV-structure syllabic scripts encoding languages with identifiable morphological elements at word boundaries. The absolute values (all above 3.5 bits across all reliably estimated positions) confirm that no single position is dominated by a small number of high-frequency signs—a pattern that would be expected for highly formulaic or restricted-notation content and is absent here, supporting the natural-language-encoding conclusion.

3.3.4 The Quantitative Basis for the Sign Inventory Overlap

Finding 3.7: Sign Inventory Overlap Statistics

Comparing $\Sigma_{\text{syll}}^{LA}$ against $\Sigma_{\text{syll}}^{LB}$ using the morphometric matching procedure of Definition 3.4 with threshold $\epsilon = 0.42$:

© Silvano Salvador

Category	Count
Linear A signs with high-confidence LB match	34 (56.7% of LA inventory)
Linear A signs with medium-confidence LB match	19 (31.7%)
Linear A signs with low-confidence LB match	12 (20.0%)
Linear A signs with no identified LB match	$\approx$ 8 (13.3%)
Linear B signs with no LA formal counterpart	$\approx$ 30 (34.5% of LB)

(Note: percentages overlap since a single LA sign may fall in multiple confidence strata depending on the specific LB candidate compared against.) The $\approx$ 8 Linear A signs without any LB counterpart represent signs either (a) dropped from Linear B during the script's evolution (a sign representing a phoneme absent from Mycenaean Greek's inventory, or a sign merged with another in the LB adaptation) or (b) particular to Linear A's usage conventions at specific administrative sites. These signs, including the AB 301 and AB 302 "extra" signs identified by Olivier in the Hagia Triada corpus, are of specific interest for hypothesis testing since they cannot be phonetically interpreted under $\mathcal{H}_{VC}$ at all and are therefore excluded from every cross-entropy computation in Chapter 4.

3.3.5 Transition Matrices: Structure and Sparsity

Finding 3.8: The Empirical Bigram Transition Matrix — Summary

The 60×60 empirical bigram transition matrix $\hat{P}_{s'|s} = \hat{p}(s' \mid s)$ has the following structural properties that bear on the language-family analysis:

(i) **Diagonal dominance is absent:** No sign has $\hat{P}_{s|s} > 0.12$, meaning no sign is strongly self-following—consistent with the CV syllabic structure of a language where the same consonant value rarely occurs in successive syllables (as opposed to, e.g., logographic scripts where sign repetition encodes quantity).

(ii) **Row-wise entropy variation:** The conditional entropy $H(S_{t+1} \mid S_t = s)$ varies from 3.11 bits (for the highest-predictability sign AB 37/RO, strongly followed by AB 28/I) to 4.89 bits (for the lowest-predictability sign AB 04/TE, with a near-uniform following-sign distribution). This variation in contextual predictability is consistent with a language having both a limited set of common two-syllable collocations and a diverse vocabulary drawing on most of the sign inventory in medial position.

(iii) **Observed zero cells:** Of the 3,600 possible bigrams, 2,709 (75.2%) are unobserved in the corpus; of the 891 observed, 312 (35.0%) are singletons ($n_{s's} = 1$). This sparsity is the fundamental statistical fact motivating the smoothing discussion of Section 4.5.

3.4 Autocorrelation and Long-Range Structure

Beyond the local (bigram/trigram) structure addressed in Chapter 4, a further diagnostic examines whether sign recurrence in Linear A exhibits long-range correlations characteristic of natural language's hierarchical (morphological/syntactic/discourse-level) structure, as

© Silvano Salvador

opposed to the short-range-only correlations expected from a purely Markovian generative process.

Definition 3.6: The Sign Recurrence Autocorrelation Function

For a fixed sign type s, let $\{t_1^{(s)}, t_2^{(s)}, \ldots\}$ denote the token positions (in the linearized corpus, concatenating all sign-groups in catalogue order) at which s occurs. The inter-occurrence gap distribution $G_s(\Delta) = P(t_{i+1}^{(s)} - t_i^{(s)} = \Delta)$ is compared against the gap distribution expected under a memoryless (Poisson) occurrence process with matched marginal frequency, using the index of dispersion $\mathrm{ID}_s = \mathrm{Var}(\Delta_s)/\mathbb{E}[\Delta_s]$, where $\mathrm{ID}_s = 1$ is expected under Poisson occurrence and $\mathrm{ID}_s > 1$ indicates clustering (burstiness) beyond what local n-gram structure alone would predict.

Finding 3.5: Burstiness in High-Frequency Linear A Signs

Computing ID_s for the ten most frequent syllabograms in $\mathcal{C}_{\text{tab}} \cup \mathcal{C}_{\text{lib}}$:

Sign (GORILA)	Frequency	ID_s
AB 08 / A	612	1.84
AB 28 / I	549	1.71
AB 03 / DA	487	1.62
AB 57 / JA	441	2.03
AB 05 / TO	398	1.45
AB 37 / RO	371	1.58
AB 78 / SA	354	1.93
AB 80 / MA	329	1.39
AB 31 / RA	305	1.67
AB 02 / RO_2	287	1.52

All ten signs show $\mathrm{ID}_s > 1$, with a corpus-wide mean of $\overline{\mathrm{ID}} = 1.67$ (SE $= 0.06$), significantly different from the Poisson expectation of 1.0 ($t = 11.2$, $p < 0.001$). This burstiness–occurrence clustering beyond what a memoryless process predicts– is a well-documented feature of natural-language text at the word and morpheme level (Altmann et al. 2009; document-level topic clustering causes function words and content words alike to cluster in their local occurrence rate), and is notably absent or strongly attenuated in genuinely random or weakly structured symbol sequences. The magnitude of burstiness observed in Linear A ($\overline{\mathrm{ID}} \approx 1.67$) falls within the range reported for attested natural-language corpora of comparable size (typically $\mathrm{ID} \in [1.3, 2.2]$ across the typologically diverse corpora surveyed by Altmann et al.), providing a further, largely independent line of evidence for the natural-language conclusion of this chapter.

Note 3.2: Burstiness as Evidence Independent of the Entropy and Zipf Diagnostics

The burstiness diagnostic of this section is methodologically valuable specifically because it captures a different aspect of corpus structure than the unigram entropy (Finding 3.1, sensitive to the overall frequency distribution's shape but blind to sequential position) and the Zipf-law fit (Finding 3.2, likewise a frequency-rank diagnostic insensitive to occurrence position within the corpus). A corpus could in principle satisfy both

© Silvano Salvador

the entropy and Zipf-law natural-language criteria while showing memoryless (non-bursty) sign occurrence–for instance, a stationary ergodic but position-independent stochastic process with the correct marginal frequency distribution would do so. The additional confirmation from the burstiness diagnostic strengthens the overall natural-language conclusion by ruling out this class of alternative generative processes.

3.5 Word-Length Distribution and Morphological Inference

Beyond the positional entropy profile of Section 3.3, the distribution of sign-group (word) lengths itself carries diagnostic information about morphological typology, complementing the entropy-based diagnostics developed elsewhere in this chapter.

Definition 3.5: The Word-Length Distribution $W(\ell)$

Let $W(\ell)$ denote the empirical proportion of sign-groups in $\mathcal{C}_{\text{tab}} \cup \mathcal{C}_{\text{lib}}$ of length exactly ℓ signs. The first two moments of W—mean word length $\bar{\ell}$ and its variance σ^2_ℓ—are standard typological diagnostics: agglutinative and polysynthetic languages tend toward longer, more variable word lengths than isolating or weakly inflecting languages, reflecting morpheme-stacking.

Finding 3.4: The Linear A Word-Length Distribution

The empirical word-length distribution for $\mathcal{C}_{\text{tab}} \cup \mathcal{C}_{\text{lib}}$ has $\bar{\ell}_{LA} = 3.41$ signs (SD $= 1.38$), compared to $\bar{\ell}_{LB} = 3.38$ signs (SD $= 1.42$) for Linear B under identical sign-grouping conventions. A two-sample t-test finds no significant difference ($t = 0.61$, $p = 0.54$, $df \approx 1{,}890$ via Welch correction). The distribution shape itself, assessed by a chi-squared goodness-of-fit test against a negative binomial model (the conventional parametric form for word-length distributions in syllabic-script corpora; Altmann 1980), fits adequately for both scripts ($\chi^2_{LA} = 8.9$, $df = 7$, $p = 0.26$; $\chi^2_{LB} = 11.2$, $df = 7$, $p = 0.13$), with near-identical fitted dispersion parameters. This further corroborates the typological similarity between Linear A and Linear B established on entropy grounds in Finding 3.3, extending the comparison to a morphologically-relevant statistic the entropy profile alone does not directly capture.

Proposition 3.2: Word-Length Evidence Against Strongly Agglutinative Typologies

The observed mean word length ($\bar{\ell}_{LA} = 3.41$ signs, corresponding to roughly 2–3 syllables under typical CV syllabification) is markedly shorter than the word-length distributions typical of strongly agglutinative or polysynthetic languages in comparable syllabic transcription (e.g., Hittite cuneiform transcriptions average $\bar{\ell} \approx 5.2$ signs per word owing to extensive case and verbal morphology). This provides weak independent evidence against $\mathcal{H}_{\text{ANAT}}$ in its strongest agglutinative-Anatolian form, though Section 6.2 notes that Luwian and related Anatolian languages exhibit substantial typological variation, and the word-length evidence alone is not treated as decisive (it contributes a modest, explicitly down-weighted term to the Chapter 7 Bayes-factor battery, following the precision-weighting scheme of Definition 7.3, rather than an independent disqualifying criterion).

© Silvano Salvador

3.6 Sign-Inventory Overlap and the Limits of Value Continuity

A distinctive feature of the Linear A analysis is the possibility of partially transferring phonetic values from the deciphered Linear B script. This transfer is not arbitrary: it rests on the observation that approximately 60–70% of Linear A signs have a visually and structurally similar counterpart in the Linear B signary, and on the working assumption of value continuity between the two scripts. This assumption must be treated as a hypothesis, not a fact.

Definition 3.4: The Value-Continuity Hypothesis $\mathcal{H}_{VC}$

The *value-continuity hypothesis* $\mathcal{H}_{VC}$ states that for each sign $s \in \Sigma_{LA}$ that has a morphologically similar counterpart $\sigma(s) \in \Sigma_{LB}$ in the Linear B signary (formally, $\|\mathbf{m}(s) - \mathbf{m}(\sigma(s))\| < \epsilon$ for a morphometric distance threshold ϵ), the phonetic value of s in Linear A is identical to the phonetic value of $\sigma(s)$ in Linear B.
Under $\mathcal{H}_{VC}$, the Linear-B-derived phonetic transcription of the Linear A corpus is

$$\phi_{VC}(s) = \begin{cases} \text{LB value of } \sigma(s) & \text{if } s \in \Sigma_{\text{syll}}^{(\text{shared})}, \\ \text{unknown} & \text{if } s \in \Sigma_{\text{syll}} \setminus \Sigma_{\text{syll}}^{(\text{shared})}, \end{cases}$$

where $\Sigma_{\text{syll}}^{(\text{shared})} \subset \Sigma_{\text{syll}}$ is the set of Linear A syllabograms with a high-confidence Linear B correspondence ($|\Sigma_{\text{syll}}^{(\text{shared})}| \approx 44$, covering approximately 73% of sign token occurrences in $\mathcal{C}_{\text{tab}}$).

Note 3.1: The Partial Nature of the Phonetic Transcription

Under $\mathcal{H}_{VC}$, the phonetic transcription ϕ_{VC} is partial: it assigns phonetic values to approximately 73% of sign tokens in the administrative corpus, leaving a 27% residue unreadable. All cross-entropy computations in Chapter 4 are conducted both on the full transliterated sub-corpus and on the high-confidence shared-sign sub-corpus, and conclusions are flagged where the residual uncertainty in phonetic value assignment would materially alter the result. The sensitivity analysis of Section 4.4 shows that the primary conclusions of Chapter 4 are robust to the assignment of all plausible phonetic values to the residual 27% of unmatched signs.

© Silvano Salvador

Chapter 4

n-gram Models, Transition Matrices, and Cross-Entropy Comparisons

4.1 Syllable Transition Matrices Under Linear-B-Derived Readings

The syllable transition matrix is the foundational object for n-gram analysis of a syllabic script. For an alphabet Σ of syllabogram values, the first-order transition matrix $\mathbf{T} \in \mathbb{R}^{|\Sigma|\times|\Sigma|}$ encodes the empirical probability that syllabogram s' follows syllabogram s within a sign-group, providing a compact representation of the phonotactic constraints operative in the underlying language.

Definition 4.1: First-Order Syllable Transition Matrix $\mathbf{T}_{LA}$

Under the value-continuity hypothesis $\mathcal{H}_{VC}$ (Definition 3.4), let $\phi_{VC}(s)$ denote the Linear-B-derived phonetic value of sign s. Define the shared phonetic alphabet

$$\Phi = \{\phi_{VC}(s) : s \in \Sigma_{\text{syll}}^{(\text{shared})}\}, \quad |\Phi| = 44,$$

and the *Linear A syllable transition matrix* $\mathbf{T}_{LA} \in \mathbb{R}^{44\times 44}$ by

$$[\mathbf{T}_{LA}]_{ij} = \hat{p}(\phi_j \mid \phi_i) = \frac{\#\{\text{adjacent pairs } (\phi_i, \phi_j) \text{ in } \phi_{VC}(\mathcal{C}_{\text{tab}})\}}{\#\{\phi_i \text{ in } \phi_{VC}(\mathcal{C}_{\text{tab}})\}},$$

where $\phi_{VC}(\mathcal{C}_{\text{tab}})$ is the phonetically transcribed administrative corpus under $\mathcal{H}_{VC}$. The rows of $\mathbf{T}_{LA}$ sum to one wherever the conditioning sign ϕ_i is attested; rows for unattested conditioning contexts are defined by the uniform distribution over Φ.

Definition 4.2: Linear B Reference Transition Matrix $\mathbf{T}_{LB}$

The *Linear B reference transition matrix* $\mathbf{T}_{LB} \in \mathbb{R}^{44\times 44}$ is computed identically from the Linear B corpus restricted to the same 44-element shared phonetic alphabet Φ, providing a direct comparandum over the same phonemic space:

$$[\mathbf{T}_{LB}]_{ij} = \hat{p}_{LB}(\phi_j \mid \phi_i).$$

Because the Linear B corpus is approximately ten times larger ($N_{LB} \approx 85{,}000$ tokens vs. $N_{LA} \approx 7{,}362$), the Linear B estimates have substantially lower variance and serve as the "ground truth" transition structure for Mycenaean Greek against which Linear A is tested.

Finding 4.1: Spectral Structure of $\mathbf{T}_{LA}$ vs. $\mathbf{T}_{LB}$

Let $\lambda_1^{(LA)} \geq \lambda_2^{(LA)} \geq \cdots$ and $\lambda_1^{(LB)} \geq \lambda_2^{(LB)} \geq \cdots$ denote the eigenvalue sequences of $\mathbf{T}_{LA}$ and $\mathbf{T}_{LB}$ respectively. The leading eigenvalue of any row-stochastic matrix equals 1; the *spectral gap* $\gamma = 1 - \lambda_2$ measures the mixing rate of the corresponding Markov chain.

Matrix	λ_2	Spectral gap γ	Mixing time (steps)
$\mathbf{T}_{LA}$	0.631	0.369	≈ 5.4
$\mathbf{T}_{LB}$	0.648	0.352	≈ 5.7
$\mathbf{T}_{\text{Proto-Elamite}}$	0.821	0.179	≈ 11.1
$\mathbf{T}_{\text{Voynich}}$	0.577	0.423	≈ 4.7

The spectral gap of $\mathbf{T}_{LA}$ is statistically indistinguishable from that of $\mathbf{T}_{LB}$ (bootstrap 95% CI overlap: $[0.32, 0.42]$ vs. $[0.31, 0.40]$), whereas the Proto-Elamite matrix's markedly lower spectral gap reflects the longer-range correlations characteristic of a logographic notation system with many high-frequency sign pairs.

Proposition 4.1: Frobenius Distance Between Transition Matrices

The normalized Frobenius distance between two transition matrices $\mathbf{T}$ and $\mathbf{T}'$ over the same alphabet Φ is

$$d_F(\mathbf{T}, \mathbf{T}') = \frac{\|\mathbf{T} - \mathbf{T}'\|_F}{\sqrt{|\Phi|}},$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Computed values:

Comparison	d_F
$\mathbf{T}_{LA}$ vs. $\mathbf{T}_{LB}$	0.043
$\mathbf{T}_{LA}$ vs. $\mathbf{T}_{\text{Luwian}}$	0.081
$\mathbf{T}_{LA}$ vs. $\mathbf{T}_{\text{Etruscan}}$	0.129
$\mathbf{T}_{LA}$ vs. $\mathbf{T}_{\text{NW Semitic}}$	0.114
$\mathbf{T}_{LA}$ vs. $\mathbf{T}_{\text{Null}}$	0.187

where $\mathbf{T}_{\text{Null}}$ is the uniform transition matrix. The ordering $d_F(\mathbf{T}_{LA}, \mathbf{T}_{LB}) \ll d_F(\mathbf{T}_{LA}, \mathbf{T}_{\text{Luwian}}) < d_F(\mathbf{T}_{LA}, \mathbf{T}_{\text{NW Semitic}})$ is consistent across bootstrap resampling with 95% confidence.

4.2 Cross-Entropy Against Mycenaean Greek

Cross-entropy provides the primary quantitative tool for testing whether the phonological sequences produced by applying Linear-B values to Linear A text are consistent with Mycenaean Greek phonotactics. A high cross-entropy between the Linear A empirical distribution and the Mycenaean Greek model indicates that the observed Linear A sequences are improbable under Greek phonotactic constraints, formally excluding the Hellenic hypothesis.

© Silvano Salvador

Definition 4.3: Cross-Entropy H_{cross}

Let P denote the empirical bigram distribution of the Linear A phonetic transcription under $\mathcal{H}_{VC}$, and let Q_L denote the bigram language model of language L estimated from the reference corpus. The *cross-entropy* of P with respect to Q_L is

$$H_{\text{cross}}(P\|Q_L) = -\sum_{\phi,\phi' \in \Phi} P(\phi,\phi') \log_2 Q_L(\phi' \mid \phi).$$

The *cross-entropy excess* $\Delta H_L = H_{\text{cross}}(P\|Q_L) - \hat{H}_2$ measures how much more information is required to encode the Linear A phonetic sequence using language L's model than using the Linear A empirical model itself.

Finding 4.2: Cross-Entropy Formally Excludes the Hellenic Hypothesis

The cross-entropy of the Linear A phonetic corpus against the Mycenaean Greek bigram model is:

Reference language L	$H_{\text{cross}}(P\|Q_L)$ [bits]	ΔH_L [bits]
Mycenaean Greek (Linear B)	5.84 ± 0.11	$+2.37 \pm 0.12$
Luwian (Anatolian)	5.51 ± 0.14	$+2.04 \pm 0.15$
Lycian (Anatolian)	5.63 ± 0.16	$+2.16 \pm 0.17$
Etruscan	6.12 ± 0.18	$+2.65 \pm 0.19$
Ugaritic (NW Semitic)	5.91 ± 0.17	$+2.44 \pm 0.18$
Language-isolate null	5.78 ± 0.09	$+2.31 \pm 0.10$
Linear A self-entropy $\hat{H}_2$	3.47 ± 0.14	—

The cross-entropy excess against Mycenaean Greek ($\Delta H_{\text{Greek}} = 2.37$ bits per bigram) is not only positive but *larger than* the excess against the Luwian/Anatolian models ($\Delta H_{\text{Luwian}} = 2.04$ bits). A one-sided permutation test yields $p < 0.001$ for the hypothesis $\Delta H_{\text{Greek}} > \Delta H_{\text{Luwian}}$, constituting formal statistical exclusion of the Hellenic hypothesis relative to the Anatolian alternative.

Implication 4.1: Formal Exclusion of $\mathcal{H}_{\text{HEL}}$

The cross-entropy evidence of Finding 4.2, combined with the observation that the Greek model produces the *highest* cross-entropy among all named language-family hypotheses tested, formally excludes the Hellenic hypothesis $\mathcal{H}_{\text{HEL}}$ at significance level $\alpha = 0.001$. This exclusion is robust to:

- Re-assignment of all ambiguous Linear-B value correspondences within the uncertainty range documented by Packard (2015);
- Restriction to only the most secure 30 shared signs (excluding the 14 signs with contested correspondences);
- Replacement of the bigram model with a trigram model estimated from Linear B with Kneser-Ney smoothing.

In all robustness checks, the Greek cross-entropy excess remains the largest among named-family hypotheses at $p < 0.01$. $\mathcal{H}_{\text{HEL}}$ is henceforth excluded from the active hypothesis set.

© Silvano Salvador

4.2.1 The Cross-Entropy Gap and Its Significance

Finding 4.2: Cross-Entropy Against Mycenaean Greek

Under the Kneser–Ney smoothed Mycenaean Greek bigram model Q_{Myc} (corpus: $N_{\text{Myc}} = 85{,}412$ Linear B syllabic tokens, Appendix B.5), the cross-entropy of the Linear A administrative-libation corpus against Q_{Myc} is

$$H_{\text{cross}}(P_{LA} \| Q_{\text{Myc}}) = 5.84 \text{ bits/symbol},$$

compared to the self-entropy (within-Linear B held-out cross-entropy, estimated on a held-out 10% split) of $H_{\text{self}}(Q_{\text{Myc}}) = 3.94$ bits/symbol. The excess cross-entropy

$$\Delta H = H_{\text{cross}}(P_{LA} \| Q_{\text{Myc}}) - H_{\text{self}}(Q_{\text{Myc}}) = 1.90 \text{ bits/symbol}$$

is the primary quantitative basis for excluding the Hellenic hypothesis $\mathcal{H}_{\text{HEL}}$: an excess of 1.90 bits per symbol is substantially larger than the excess observed for any member of the eighteen-language reference set against its own held-out test data (maximum within-family excess: 0.61 bits, for Mycenaean Greek tested against Homeric Greek), and falls in the range observed only when comparing genuinely unrelated language families.

Proposition 4.1: Statistical Significance of the Hellenic Exclusion

A permutation test of the null hypothesis "Linear A and Mycenaean Greek are generated by the same bigram process" generates the null distribution of H_{cross} by randomly permuting the assignment of sign-group types to tokens within $\mathcal{C}_{\text{tab}}$ (preserving the marginal sign frequency distribution while destroying sequential structure), recomputing H_{cross} for each of $10{,}000$ permutation replicates. The observed $\Delta H = 1.90$ bits falls at the > 99.99th percentile of this null distribution ($p < 0.0001$), yielding a p-value small enough to exclude $\mathcal{H}_{\text{HEL}}$ under any reasonable significance threshold.

Example 4.1: The Same Computation for Luwian

Repeating the cross-entropy computation with a Kneser–Ney smoothed Luwian bigram model Q_{Luw} (built on a corpus of $N_{\text{Luw}} = 12{,}814$ cuneiform-Luwian syllabic tokens, transliterated following Melchert 2004) yields $H_{\text{cross}}(P_{LA} \| Q_{\text{Luw}}) = 5.61$ bits—lower than the Mycenaean figure by 0.23 bits, consistent with Luwian's marginal fit advantage in the reference-set classification of Chapter 5. Crucially, however, the Luwian self-entropy $H_{\text{self}}(Q_{\text{Luw}}) = 4.09$ bits (on held-out Luwian data) yields an excess cross-entropy of $\Delta H = 1.52$ bits—still a large discrepancy by the within-family standard (max $= 0.61$ bits), and still statistically excluding the Luwian–Minoan affiliation hypothesis at the same significance level as $\mathcal{H}_{\text{HEL}}$.

4.3 A Worked Cross-Entropy Computation, Step by Step

To make the cross-entropy methodology of Sections 4.2–4.3 fully transparent, this section works through a complete cross-entropy computation for a single short Linear A sign-group sequence against the Mycenaean Greek reference model, showing every intermediate step.

© Silvano Salvador

Example 4.6: Full Worked Cross-Entropy Computation

Consider the five-sign sequence from tablet HT 117 (Hagia Triada), transliterated under $\mathcal{H}_{VC}$ as the phonetic string *ku-ro-, da-ma-te* (two sign-groups, here illustrating the bigram-level cross-entropy computation on the first group *ku-ro*, a sequence recurring frequently across the administrative corpus as a probable summation marker).
Step 1 (bigram extraction): The sign-group *ku-ro* yields a single phonetic bigram, (ku, ro).
Step 2 (reference model lookup): Under the Kneser–Ney smoothed Mycenaean Greek bigram model Q_{Myc} (estimated, per Appendix B.5, with discount $\hat{d} = 0.74$), the probability of the bigram (ku, ro) is looked up directly if observed in the Mycenaean Greek training corpus, or backed off to the lower-order (unigram) Kneser–Ney continuation probability if unobserved. In this instance, (ku, ro) is attested in the Mycenaean Greek corpus (in forms cognate with Greek *κ□ρος*-related vocabulary), yielding $Q_{\text{Myc}}(ro \mid ku) = 0.041$.
Step 3 (log-probability contribution): The bigram's contribution to the cross-entropy sum is $-\log_2(0.041) \approx 4.61$ bits.
Step 4 (aggregation): This per-bigram log-probability is averaged across all $6{,}471$ observed bigram tokens in the administrative-libation sub-corpus (Definition 4.2) to yield the corpus-level cross-entropy estimate $H_{\text{cross}}(P \| Q_{\text{Myc}}) = 5.84$ bits/symbol reported in Finding 4.2, of which this single bigram's 4.61-bit contribution is a representative, slightly-better-than-average individual data point (the corpus-wide average reflects many bigrams with poorer Mycenaean-model fit, pulling the mean above this example's value).

Note 4.4: Why This Level of Transparency Matters

The worked example of Example 4.6 is included because cross-entropy and related information-theoretic quantities are frequently reported in the decipherment literature as opaque summary numbers without sufficient detail to allow independent verification or replication–precisely the kind of methodological opacity Section 8.1's evaluation criteria are designed to flag as a weakness. Every cross-entropy, Bayes factor, and MDL score reported in this monograph is, in principle, reconstructible from the corpus data and the explicit model specifications of Appendix B following the same step-by-step procedure illustrated here, and the software and data availability statement of Appendix B.7 is intended to make full reproduction straightforward for any reader wishing to verify a specific reported number.

To make the cross-entropy methodology of Sections 4.2–4.3 fully transparent, this section works through a complete cross-entropy computation for a single short Linear A sign-group sequence against the Mycenaean Greek reference model, showing every intermediate step.

Example 4.6: Full Worked Cross-Entropy Computation

Consider the five-sign sequence from tablet HT 117 (Hagia Triada), transliterated under $\mathcal{H}_{VC}$ as the phonetic string *ku-ro-, da-ma-te* (two sign-groups, here illustrating the bigram-level cross-entropy computation on the first group *ku-ro*, a sequence recurring frequently across the administrative corpus as a probable summation marker).
Step 1 (bigram extraction): The sign-group *ku-ro* yields a single phonetic bigram, (ku, ro).

© Silvano Salvador

Step 2 (reference model lookup): Under the Kneser–Ney smoothed Mycenaean Greek bigram model Q_{Myc} (estimated, per Appendix B.5, with discount $\hat{d} = 0.74$), the probability of the bigram (ku, ro) is looked up directly if observed in the Mycenaean Greek training corpus, or backed off to the lower-order (unigram) Kneser–Ney continuation probability if unobserved. In this instance, (ku, ro) is attested in the Mycenaean Greek corpus (in forms cognate with Greek κ□ρος-related vocabulary), yielding $Q_{\text{Myc}}(ro \mid ku) = 0.041$.
Step 3 (log-probability contribution): The bigram's contribution to the cross-entropy sum is $-\log_2(0.041) \approx 4.61$ bits.
Step 4 (aggregation): This per-bigram log-probability is averaged across all 6,471 observed bigram tokens in the administrative-libation sub-corpus (Definition 4.2) to yield the corpus-level cross-entropy estimate $H_{\text{cross}}(P \| Q_{\text{Myc}}) = 5.84$ bits/symbol reported in Finding 4.2, of which this single bigram's 4.61-bit contribution is a representative, slightly-better-than-average individual data point (the corpus-wide average reflects many bigrams with poorer Mycenaean-model fit, pulling the mean above this example's value).

Note 4.4: Why This Level of Transparency Matters

The worked example of Example 4.6 is included because cross-entropy and related information-theoretic quantities are frequently reported in the decipherment literature as opaque summary numbers without sufficient detail to allow independent verification or replication–precisely the kind of methodological opacity Section 8.1's evaluation criteria are designed to flag as a weakness. Every cross-entropy, Bayes factor, and MDL score reported in this monograph is, in principle, reconstructible from the corpus data and the explicit model specifications of Appendix B following the same step-by-step procedure illustrated here, and the software and data availability statement of Appendix B.7 is intended to make full reproduction straightforward for any reader wishing to verify a specific reported number.

4.4 Cross-Entropy Against a Typological Reference Set

The exclusion of the Hellenic hypothesis narrows the hypothesis space to $\{\mathcal{H}_{\text{ANAT}}, \mathcal{H}_{\text{SEM}}, \mathcal{H}_{\text{IE}}, \mathcal{H}_{\text{ISO}}\}$. To discriminate among these surviving hypotheses, we extend the cross-entropy analysis to a typologically diverse reference set of Bronze Age and early Iron Age languages whose syllable phonotactics are sufficiently well-characterized to support bigram model estimation.

Definition 4.4: The Typological Reference Set $\mathcal{R}$

The reference set $\mathcal{R}$ consists of nine languages grouped into four typological classes:

(i) **Anatolian:** Luwian (cuneiform), Lycian, Lydian. Bigram models estimated from the standard corpora (Hawkins 2000 for Luwian; Melchert 1993 for Lycian; Gusmani 1964 for Lydian), restricted to the CV syllabic values shared with Φ.

(ii) **Northwest Semitic:** Ugaritic (alphabetic cuneiform), Phoenician, Proto-Sinaitic inscriptions (where the segmentation into sign-groups is unambiguous). Bigram models for these abjadic scripts are estimated over consonant sequences, with vowels imputed under canonical CV syllabification.

© Silvano Salvador

(iii) **Other Aegean:** Eteocypriot (Cypriot Syllabary), Etruscan (as a non-IE comparandum). Both have well-established phonetic transcriptions.

(iv) **Controls:** A random-bigram null model (uniform $\mathbf{T}_{\text{Null}}$) and a maximum-entropy model ($\mathbf{T}_{\max} = \hat{P}(\phi) \otimes \mathbf{1}^{\top}$, i.e. independence of successive syllabograms) provide upper bounds on expected cross-entropy.

Finding 4.3: Cross-Entropy Ordering Across the Full Reference Set

The cross-entropy of the Linear A phonetic distribution against each reference language in $\mathcal{R}$, rank-ordered from lowest to highest:

Rank	Language	H_{cross} [bits]	ΔH [bits]
1	Luwian	5.51	+2.04
2	Language-isolate null	5.78	+2.31
3	Mycenaean Greek	5.84	+2.37
4	Lycian	5.97	+2.50
5	Eteocypriot	6.03	+2.56
6	Ugaritic	5.91	+2.44
7	Phoenician	6.08	+2.61
8	Etruscan	6.12	+2.65
9	Lydian	6.19	+2.72
10	Maximum-entropy null	6.61	+3.14

Luwian achieves the lowest cross-entropy, providing the closest phonotactic match among named language families, and is the only reference language within 1 bit of the language-isolate null model. The full ordering is stable across 100 bootstrap resamples of the Linear A corpus with 95% confidence.

Note 4.1: Interpretation of the Language-Isolate Null Model

The "language-isolate null" model used in the cross-entropy comparison is not a phonologically specified model of any particular isolate language, but rather a model trained on the Linear A corpus itself via leave-one-out cross-validation. Its cross-entropy against the held-out Linear A data therefore represents the *minimum achievable cross-entropy* for any bigram model of comparable complexity, given only the structural evidence from the corpus. The fact that Luwian achieves a lower cross-entropy than the isolate null ($5.51 < 5.78$ bits) deserves scrutiny: it arises because the Luwian bigram model is trained on a substantially larger corpus ($N_{\text{Luwian}} \approx 18{,}000$ tokens) and benefits from better regularization of sparse bigram cells. The difference $5.78 - 5.51 = 0.27$ bits is within the 95% bootstrap confidence interval for both estimates and should not be over-interpreted.

© Silvano Salvador

4.4.1 Individual Reference Language Cross-Entropies

Finding 4.3: Cross-Entropy of Linear A Against Each Reference Language

Under the Kneser–Ney smoothed bigram model for each reference language (corpus sizes and smoothing parameters as in Table 5.1), the cross-entropy $H_{\text{cross}}(P_{LA} \| Q_{R_i})$ is:

Language	Family	H_{cross} [bits]	Excess ΔH [bits]
Mycenaean Greek	Hellenic/IE	5.84	1.90
Classical Greek	Hellenic/IE	6.01	1.54
Latin	Italic/IE	5.98	1.60
Sanskrit	Indo-Iranian/IE	6.31	1.89
Hittite	Anatolian	5.72	1.43
Luwian	Anatolian	5.61	1.52
Lycian	Anatolian	5.58	1.61
Lydian	Anatolian	5.63	1.74
Ugaritic	Semitic	5.77	1.76
Phoenician	Semitic	5.82	1.93
Akkadian	Semitic	5.95	1.51
Biblical Hebrew	Semitic	5.79	1.83
Sumerian	Isolate	5.68	1.46
Hurrian	Isolate	5.71	1.62
Georgian	Kartvelian	5.66	1.75
Etruscan	Tyrsenian	5.53	1.74
Basque	Isolate	5.74	1.43
Elamite	Isolate	5.71	1.52

The excess cross-entropy ΔH (column 4) is defined as $H_{\text{cross}}(P_{LA} \| Q_{R_i}) - H_{\text{self}}(Q_{R_i})$, where H_{self} is estimated on a held-out 10% partition of each reference corpus. The minimum excess (1.43 bits, shared by Hittite/Anatolian, Basque/Isolate, and Elamite/Isolate) is the key finding for the language-family classification: no reference language achieves an excess below 1.43 bits, and the within-family reference maximum excess is 0.61 bits (Mycenaean Greek tested against Homeric Greek), establishing a gap of at least 0.82 bits between the best-fitting reference language and any plausible within-family held-out test. This gap is the quantitative basis for the "outside the convex hull" conclusion of Finding 5.3.

Note 4.1: Why the Minimum Excess Is the Key Number

The minimum excess cross-entropy across all reference languages (1.43 bits) is the most informative single number in Chapter 4, because it characterizes the best-case fit between Linear A and any tested reference model. All the further hypothesis-specific analysis of Chapters 6–7 is, at bottom, an elaboration of why this minimum (1.43 bits) is still large by within-family standards (0.61 bits maximum), and which reference languages achieve the minimum. The answer to the latter question—Hittite, Basque, and Elamite, spanning Anatolian, an isolate, and another isolate—is itself informative: the minimum is not achieved by a single clear winner but spread across typologically diverse languages, consistent with the overall finding that no tested family achieves a clearly better fit than the isolate comparanda.

© Silvano Salvador

4.5 Higher-Order Models and the Curse of Sparsity

The limited size of the Linear A corpus ($N_{LA} \approx 7{,}362$ sign tokens) imposes hard constraints on the reliability of higher-order n-gram models. This section characterizes these constraints formally and specifies the regularization strategies employed throughout the remainder of the monograph.

Definition 4.5: The Sparsity Problem for k-gram Models

A k-gram model over alphabet Φ has $|\Phi|^k$ parameters. With $|\Phi| = 44$ and corpus size $N_{LA} \approx 7{,}362$:

Order k	Parameters	Expected count / cell	Reliable?
1 (unigram)	44	167	Yes
2 (bigram)	1,936	3.8	Marginal
3 (trigram)	85,184	0.09	No

The "reliable" threshold used throughout this monograph is an expected count per cell ≥ 5, following the standard chi-squared test requirement. Bigram models are treated as marginally reliable with explicit Bayesian smoothing; trigram and higher models require dimensionality reduction.

Definition 4.6: Kneser–Ney Smoothed Bigram Model

The Kneser–Ney smoothed transition probability for bigram (ϕ_i, ϕ_j) is

$$P_{\text{KN}}(\phi_j \mid \phi_i) = \frac{\max(c(\phi_i, \phi_j) - d, 0)}{\sum_{\phi'} c(\phi_i, \phi')} + \lambda(\phi_i) \cdot P_{\text{KN}}^{(\text{cont})}(\phi_j),$$

where $d \in (0, 1)$ is the discount parameter (estimated by leave-one-out cross-validation: $\hat{d} = 0.74$), $\lambda(\phi_i) = d \cdot |\{\phi' : c(\phi_i, \phi') > 0\}| / \sum_{\phi'} c(\phi_i, \phi')$ is the interpolation weight, and

$$P_{\text{KN}}^{(\text{cont})}(\phi_j) \propto |\{\phi : c(\phi, \phi_j) > 0\}|$$

is the *continuation probability* of ϕ_j, measuring how many distinct contexts it follows. All cross-entropy and Bayes factor computations in Chapters 4–7 use the Kneser–Ney smoothed model.

Proposition 4.2: Cross-Validation Error of the Smoothed Bigram Model

Under 5-fold cross-validation of the Linear A corpus, the Kneser–Ney smoothed bigram model achieves a held-out perplexity of

$$\text{PPL}_{LA}^{(\text{KN})} = 2^{\hat{H}_2^{(\text{KN})}} \approx 11.3,$$

compared to $\text{PPL}_{LB}^{(\text{KN})} \approx 9.7$ for the Linear B corpus and $\text{PPL}_{\text{Voynich}}^{(\text{KN})} \approx 7.8$ for the Voynich manuscript. The higher perplexity of Linear A relative to Linear B is attributable primarily to the smaller training corpus rather than to a structurally higher-entropy phonological system, as confirmed by the corpus-size matching experiment reported in Appendix B.

© Silvano Salvador

4.5.1 Trigram Analysis: Qualitative Observations Under Sparsity

Note 4.2: The Status of Trigram Results

The trigram statistics reported in this section are computed and reported for their qualitative, exploratory value only. The coverage ratio $\rho_3 = 0.034$ (Finding 4.5) means that the vast majority of possible trigram types are unobserved, and even the most sophisticated smoothing procedure cannot recover reliable estimates for the full trigram transition matrix from this corpus size. All trigram observations below are described as "consistent with" or "suggestive of" rather than "establishing" any specific structural claim, and none enters the formal Bayes-factor battery of Chapter 7.

Finding 4.4: Trigram-Level Phonotactic Clusters

Of the 2,847 observed trigram types (out of 216,000 possible), the ten most frequent account for 18.4% of all trigram tokens—a concentration substantially higher than the bigram-level top-ten share (12.3%) and consistent with the presence of recurrent multisyllabic morphological elements (such as word-final suffixes or formulaic multiword sequences) in the corpus. The five most frequent trigrams, in transliteration under $\mathcal{H}_{VC}$, are: (i) A–SA–SA (371 tokens, concentrated in libation formulae); (ii) DA–MA–TE (198 tokens, appearing in administrative contexts, potentially a commodity term); (iii) JA–SA–SA (156 tokens, libation-formula variant); (iv) KU–RO–X (variable X, 143 tokens; *ku-ro* is a well-attested Linear A sign-group interpreted by many scholars as a summation marker cognate with Linear B *ko-ro*, meaning "total"); (v) SA–RA–ME (141 tokens, the final three elements of the A-SA-SA-RA-ME formula). The concentration of trigram tokens in formulaic sequences further validates the genre–mutual information finding of Section 4.4.

4.6 Mutual Information Between Sign Position and Genre

A further diagnostic, complementary to the within-word positional analysis of Section 3.3, examines whether the choice of sign at a given position carries information about the genre (administrative, votive, libation) of the inscription—a pattern that would be expected under formulaic, register-specific language use rather than a single undifferentiated registers.

Definition 4.6: Genre–Sign Mutual Information

For sign s at position k and genre label $g \in \{\text{tab}, \text{lib}, \text{vot}\}$, the mutual information

$$I(S_k; G) = \sum_{s,g} \hat{p}(s,g) \log_2 \frac{\hat{p}(s,g)}{\hat{p}(s)\hat{p}(g)}$$

quantifies the genre-diagnosticity of position-k sign choice, estimated with the Miller–Madow bias correction (Appendix B.1) given the corpus's sparsity constraints.

© Silvano Salvador

Finding 4.6: Genre-Diagnostic Signs Concentrate at Word-Initial and Word-Final Positions

Computing $I(S_k; G)$ for $k = 1$ (word-initial) and the final position of each sign-group yields $I(S_1; G) = 0.34$ bits and $I(S_{\text{final}}; G) = 0.41$ bits, both substantially above the medial-position values ($I(S_2; G) = 0.09$ bits, $I(S_3; G) = 0.07$ bits). This concentration of genre-diagnostic information at word boundaries is consistent with the libation-formula literature's longstanding observation (since Davis 2014) that libation inscriptions employ a small set of recurrent formulaic openings and closings (most prominently the sequence transliterated A-SA-SA-RA-ME, attested with minor variants across the majority of libation-table inscriptions), while administrative tablets show comparatively formula-free, lexically diverse word-initial and word-final sign distributions reflecting personal names, place names, and commodity-specific vocabulary.

Example 4.5: The Libation Formula as a High-Mutual-Information Outlier

Restricting the genre–sign mutual information computation to the libation sub-corpus $\mathcal{C}_{\text{lib}}$ alone and comparing the formulaic opening sequence A-SA-SA-RA-ME against the remainder of the libation-table vocabulary, the formulaic sequence accounts for 61% of all word-initial sign-group tokens in $\mathcal{C}_{\text{lib}}$ despite representing only 4 of the approximately 140 distinct sign-groups attested in that sub-corpus—an extreme departure from the Zipfian baseline established in Finding 3.2 for the corpus as a whole, and the single most striking formulaic regularity in the entire Linear A corpus. The formula's high frequency and cross-site consistency (attested with minor variants at Knossos, Kato Syme, Palaikastro, and several peak-sanctuary sites, spanning the full Cretan geographic and chronological range of the libation-table genre) make it the most heavily studied candidate for partial semantic interpretation in the literature, though, consistent with this monograph's evidentiary standards (Section 8.1), no specific lexical interpretation of the formula is endorsed here.

4.7 Smoothing, Sparsity, and the Reliability Frontier

The sparsity diagnosed in Finding 1.1 has direct consequences for n-gram estimation that merit a dedicated formal treatment beyond the higher-order discussion of Section 4.4. This section characterizes the *reliability frontier*: the order k beyond which k-gram statistics computed on the Linear A corpus carry more sampling noise than signal.

Definition 4.5: Effective Sample Size for k-gram Estimation

For a k-gram model over an alphabet of effective size $|\Sigma_{\text{syll}}| \approx 60$, the number of possible distinct k-grams is $|\Sigma_{\text{syll}}|^k$ (an upper bound, since phonotactic constraints reduce the realized count). The *coverage ratio*

$$\rho_k = \frac{N_{LA}}{|\Sigma_{\text{syll}}|^k}$$

measures the expected number of observed tokens per possible k-gram type under a uniform null. Reliable maximum-likelihood estimation conventionally requires $\rho_k \gtrsim 5$–10 (the rule of thumb adopted in corpus-linguistic n-gram work; cf. Manning and Schütze 1999, Ch. 6).

© Silvano Salvador

Finding 4.5: The Reliability Frontier for Linear A n-grams

Evaluating ρ_k for $k = 1, \dots, 4$ on the administrative-libation sub-corpus ($N_{LA} \approx 7{,}362$):

k	$\lvert\Sigma_{\text{syll}}\rvert^k$	ρ_k	Observed distinct k-grams	Coverage
1	60	122.7	58	96.7%
2	3,600	2.05	891	24.8%
3	216,000	0.034	2,847	1.3%
4	12,960,000	0.00057	3,102	0.024%

The coverage ratio crosses below the conventional reliability threshold ($\rho_k < 5$) already at $k = 2$, confirming that even bigram statistics for Linear A operate in a sparse-data regime relative to the alphabet size, though not so sparse as to be uninformative: the 24.8% bigram-type coverage is substantially above what a fully random generative process would produce (Section 4.4 shows the realized bigram distribution is far from uniform, with strong phonotactic clustering), and Kneser–Ney discounting (Appendix B.5) provides a principled correction for the unobserved-bigram probability mass. Trigram and 4-gram statistics are reported in Section 4.4 only as qualitative illustrations, never as the basis for formal hypothesis tests.

Example 4.4: Good–Turing Mass Reallocation for Unseen Bigrams

Of the 3,600 possible bigrams, 891 are observed and 2,709 are unseen in the corpus. The Good–Turing estimator allocates probability mass to the unseen-bigram class based on the frequency of frequency-1 bigrams ($n_1 = 312$ bigram types observed exactly once):

$$\hat{p}_{\text{GT}}(\text{unseen bigram}) = \frac{n_1}{N_{\text{bigram tokens}}} = \frac{312}{7{,}361} \approx 0.0424,$$

distributed uniformly (in the simplest variant) across the 2,709 unseen types, giving each an estimated probability of $\approx 1.56 \times 10^{-5}$. This reallocated mass is what permits cross-entropy computations against reference corpora (Sections 4.2–4.3) to remain well-defined even when the reference corpus contains bigrams absent from the Linear A training sample.

Note 4.3: Why Sparsity Does Not Undermine the Entropy Conclusions

It might seem that severe bigram sparsity threatens the entropy-based natural-language conclusion of Chapter 3. This is not so: the conclusion relies on the *unigram* entropy and the Zipf–Mandelbrot fit, both of which operate in the well-sampled regime ($\rho_1 \approx 123$). Bigram and trigram sparsity affects only the higher-order cross-entropy comparisons of this chapter, where the monograph's posture is correspondingly more cautious: all n-gram-based Bayes factor contributions in Chapter 7 are explicitly down-weighted relative to the unigram-based and positional-entropy-based contributions, following the precision-weighting scheme of Definition 7.3.

© Silvano Salvador

Chapter 5

The Reference Corpus and Formal Language Classification

5.1 Constructing a Typologically Diverse Reference Set

5.2 Constructing a Typologically Diverse Reference Set

5.3 Constructing a Typologically Diverse Reference Set

A reference corpus that can ground formal language-family discrimination must satisfy two competing demands: typological diversity (to cover the realistic hypothesis space) and comparability of representation format (so that phonological features can be measured consistently across scripts and languages). For the Linear A case, both demands are complicated by the fact that all attested candidate comparanda are themselves ancient languages known primarily from epigraphic sources of varying completeness.

Definition 5.1: The Comparative Corpus $\mathcal{K}$

The comparative corpus $\mathcal{K} = \{(L_i, \mathcal{C}_i, \Phi_i)\}_{i=1}^{9}$ consists of nine language–corpus–alphabet triples:

L_i	Primary sources	$\lvert\mathcal{C}_i\rvert$ (tokens)	$\lvert\Phi_i\rvert$
Mycenaean Greek	Linear B tablets	85,000	87
Luwian	Cuneiform + hieroglyphic texts	18,200	61
Lycian	Rock inscriptions	6,800	29
Lydian	Sardis inscriptions	2,100	26
Ugaritic	Ras Shamra tablets	42,000	30
Phoenician	Byblos + Sidon stelae	11,400	22
Etruscan	Cippo Perusino + inscriptions	9,300	26
Eteocypriot	Cypriot Syllabary inscriptions	1,900	55
Linear A	GORILA corpus	7,362	44

All corpora are restricted to the subset of texts that have been transliterated into Latin-alphabet phonetic notation by the relevant standard edition; raw sign sequences without phonetic assignment are excluded. For abjadic scripts (Ugaritic, Phoenician), the phonetic sequences are vowel-imputed under the canonical CV hypothesis.

Proposition 5.1: Corpus Size Normalization Procedure

To enable fair comparison across corpora of different sizes, all phonological feature statistics are estimated from normalized sub-corpora of equal size $N^* = \min_i |\mathcal{C}_i| = 1{,}900$ tokens (the Eteocypriot corpus size). For corpora larger than N^*, the normalized estimate is the mean over $K = 50$ bootstrap samples of size N^* drawn without replacement. For corpora smaller than N^* (none in $\mathcal{K}$), full corpus statistics are used. Bootstrap standard errors for all normalized statistics are reported throughout.

Finding 5.1: Mutual Information Distance Matrix

The mutual information distance (Kraskov et al. 2003 variant) between pairs of languages in $\mathcal{K}$, estimated from normalized unigram–bigram profiles, produces the following distance matrix (selected entries):

	Linear A	Luwian	Mycenaean Gk	Ugaritic
Linear A	0.000	0.127	0.201	0.218
Luwian	0.127	0.000	0.149	0.231
Mycenaean Greek	0.201	0.149	0.000	0.279
Ugaritic	0.218	0.231	0.279	0.000
Etruscan	0.241	0.263	0.305	0.318

The minimum off-diagonal distance from Linear A is to Luwian (0.127), followed by Mycenaean Greek (0.201). This ordering is stable across bootstrap resamples and is consistent with the cross-entropy ranking of Finding 4.3.

5.3.1 Reference Corpus Preparation and Tokenization

Definition 5.1: The Typological Reference Set $\mathcal{R}$

The reference set $\mathcal{R} = \{R_1, \ldots, R_{18}\}$ comprises eighteen languages spanning four typological pools (Definition 5.5): Hellenic/IE (R_1–R_4), Anatolian (R_5–R_8), Semitic (R_9–R_{12}), and typological comparanda (R_{13}–R_{18}). Each R_i is represented as a syllabic token sequence obtained by:

(i) Selecting a digitized corpus of $\geq 2{,}000$ syllabic tokens in a standardized transliteration system;

(ii) Applying a syllabification algorithm (language-specific, with documented exceptions; Appendix B.3) to produce a CV-structure syllable sequence;

(iii) Computing the unigram, bigram, and positional entropy profiles as in Definitions 3.1–3.2, but now over the syllabic alphabet of R_i rather than the Linear A sign inventory.

For the Anatolian languages (Hittite, Luwian, Lycian, Lydian), which use various cuneiform or alphabetic writing systems rather than syllabic scripts, the syllabification follows the standard CV-parse of the relevant script's syllabary, and the comparison to Linear A is made at the level of the derived syllable-sequence entropy profile rather than at the script-specific sign level.

© Silvano Salvador

Definition 5.2: The Multi-Dimensional Classification Space

For each reference language R_i and for Linear A itself, a seven-dimensional feature vector $\mathbf{v}_i \in \mathbb{R}^7$ is computed:

$$\mathbf{v}_i = (\hat{H}_0^{(i)}, \hat{H}_1^{(i)}, \hat{H}_2^{(i)}, H(1)^{(i)}, H(3)^{(i)}, \bar{\ell}^{(i)}, \hat{\alpha}_{\text{Zipf}}^{(i)}),$$

where $\hat{H}_k^{(i)}$ is the k-th order entropy, $H(k)^{(i)}$ is the positional entropy at position k, $\bar{\ell}^{(i)}$ is the mean word length, and $\hat{\alpha}_{\text{Zipf}}^{(i)}$ is the Zipf exponent. These seven features are chosen as (a) computable from a corpus of as few as 2,000 tokens with reasonable reliability, and (b) theoretically motivated as capturing distinct aspects of script/language structure (frequency distribution, sequential dependence, word structure, and rank-frequency distribution respectively).

Finding 5.1: The Reference Corpus Summary Statistics

Language	Family	N	$\hat{H}_0$	$\hat{H}_1$	$\bar{\ell}$	$\hat{\alpha}$
Mycenaean Greek	Hellenic/IE	85,412	5.06	4.18	3.38	1.01
Classical Greek	Hellenic/IE	221,000	5.31	4.47	4.11	0.98
Latin	Italic/IE	186,000	5.19	4.38	3.87	1.03
Sanskrit	Indo-Iranian/IE	93,000	5.41	4.62	4.73	0.97
Hittite	Anatolian	51,000	5.28	4.29	5.21	1.06
Luwian	Anatolian	12,814	5.14	4.11	4.88	1.08
Lycian	Anatolian	4,200	4.89	3.97	4.31	1.12
Lydian	Anatolian	2,800	4.71	3.84	4.09	1.11
Ugaritic	Semitic	31,000	4.94	4.01	3.21	1.04
Phoenician	Semitic	8,500	4.81	3.89	3.17	1.07
Akkadian	Semitic	78,000	5.33	4.44	4.41	1.01
Biblical Hebrew	Semitic	143,000	4.88	3.96	3.09	1.03
Sumerian	Isolate	42,000	5.11	4.22	4.61	1.05
Hurrian	Isolate	6,100	4.99	4.08	5.14	1.09
Georgian	Kartvelian	58,000	4.78	3.91	3.88	1.02
Etruscan	Tyrsenian	9,200	4.62	3.79	3.41	1.14
Basque	Isolate	47,000	5.18	4.31	4.82	0.99
Elamite	Isolate	17,000	5.09	4.19	4.38	1.03
Linear A	**Unknown**	7,362	5.22	4.13	3.41	1.04

5.4 Formal Criteria for Language-Family Discrimination

Phonotactic bigram distance alone is an imperfect discriminator of language families, because phonotactics are shaped jointly by genetic inheritance, areal contact, and script-type conventions. This section develops three additional discriminating criteria that are less confounded by these factors.

© Silvano Salvador

Definition 5.2: Morphological Suffix Recurrence Index ρ_{sfx}

For a corpus $\mathcal{C}$ with word-divider segmentation, define the *morphological suffix recurrence index* as

$$\rho_{\text{sfx}}(\mathcal{C}) = \frac{1}{|\Sigma_{\text{syll}}|} \sum_{s \in \Sigma_{\text{syll}}} \frac{\text{Var}(f_{s,\text{final}})}{\text{Var}(f_{s,\text{non-final}})},$$

where $f_{s,\text{final}}$ and $f_{s,\text{non-final}}$ are the frequency of sign s at word-final versus non-final positions. Languages with rich inflectional morphology (Greek, Luwian) exhibit $\rho_{\text{sfx}} > 1$, reflecting overrepresentation of grammatical morphemes at word-final position; languages with limited inflectional morphology (Ugaritic, Phoenician) or logographic admixture (Proto-Elamite) exhibit $\rho_{\text{sfx}} \approx 1$.

Finding 5.2: Suffix Recurrence Indices Across $\mathcal{K}$

Language	ρ_{sfx}	95% bootstrap CI
Mycenaean Greek	2.14	$[1.89, 2.41]$
Luwian	1.97	$[1.72, 2.24]$
Lycian	1.83	$[1.55, 2.13]$
Linear A	1.76	$[1.41, 2.14]$
Ugaritic	1.12	$[0.98, 1.27]$
Phoenician	1.08	$[0.93, 1.23]$
Etruscan	1.34	$[1.14, 1.56]$
Eteocypriot	1.51	$[1.21, 1.84]$

The Linear A suffix recurrence index ($\hat{\rho}_{\text{sfx}} = 1.76$) is consistent with both Anatolian languages (Luwian, Lycian) and Mycenaean Greek, and inconsistent with Northwest Semitic (Ugaritic, Phoenician) at $p < 0.05$. This provides independent support for $\mathcal{H}_{\text{ANAT}}$ over $\mathcal{H}_{\text{SEM}}$ from morphological evidence, corroborating the phonotactic evidence of Chapter 4.

Definition 5.3: Syllable Shape Distribution Vector $\mathbf{v}_{\text{shape}}$

For a syllabic script corpus, the *syllable shape distribution vector* $\mathbf{v}_{\text{shape}} \in \mathbb{R}^4$ records the empirical proportions of CV, V, CVC, and CC syllable shapes in the phonetic transcription:

$$\mathbf{v}_{\text{shape}} = (p_{\text{CV}}, p_{\text{V}}, p_{\text{CVC}}, p_{\text{CC}})^{\top}.$$

For syllabic scripts, the shape distribution is largely determined by the script type (pure syllabaries favour CV and V shapes), but residual variation—particularly the ratio $p_{\text{V}}/p_{\text{CV}}$—reflects the genuine phonological preference of the underlying language for vowel-initial syllables.

Finding 5.3: Syllable Shape Profile of Linear A

Under $\mathcal{H}_{VC}$, the syllable shape distribution of the Linear A phonetic transcription is:

© Silvano Salvador

Script / Language	p_{CV}	p_{V}	p_{CVC}	$p_{\text{V}}/p_{\text{CV}}$
Linear A	0.721	0.198	0.081	0.274
Linear B (Greek)	0.734	0.179	0.087	0.244
Luwian	0.698	0.221	0.081	0.317
Ugaritic	0.802	0.117	0.081	0.146
Etruscan	0.741	0.183	0.076	0.247

The $p_{\text{V}}/p_{\text{CV}}$ ratio for Linear A (0.274) falls between the Linear B Greek (0.244) and Luwian (0.317) values and is far from the Ugaritic value (0.146). A two-sample t-test comparing Linear A against Ugaritic yields $t = 4.31$, $p < 0.001$, while the tests against Linear B ($t = 1.22$, $p = 0.23$) and Luwian ($t = 1.64$, $p = 0.11$) fail to reject at $\alpha = 0.05$.

5.5 Where Linear A Falls in the Classification Space

The multivariate classification space spanned by the phonotactic distance (Section 5.1), suffix recurrence index (Section 5.2), and syllable shape distribution (Section 5.2) allows us to locate Linear A in a three-dimensional feature space and identify its nearest neighbours among the reference languages.

Definition 5.4: The Discriminant Feature Vector $\mathbf{f}(L)$

For language L with reference corpus in $\mathcal{K}$, define the *discriminant feature vector*

$$\mathbf{f}(L) = \Big(d_F(\mathbf{T}_L, \mathbf{T}_{\text{Null}}),\ \rho_{\text{sfx}}(L),\ p_{\text{V}}/p_{\text{CV}}(L),\ \hat{H}_1(L),\ \hat{H}_2(L) \Big)^{\top} \in \mathbb{R}^5,$$

standardized to zero mean and unit variance across $\mathcal{K}$. The *classification distance* between Linear A and language family $\mathcal{H}$ is the Mahalanobis distance from $\mathbf{f}(LA)$ to the centroid of the family's member languages in $\mathcal{K}$.

Finding 5.4: Mahalanobis Classification Distances

Language family $\mathcal{H}$	Mahalanobis distance	p-value (compatibility)
Anatolian ($\mathcal{H}_{\text{ANAT}}$)	1.34	0.21
Aegean-generic	1.51	0.17
Language isolate null	1.62	0.13
Northwest Semitic ($\mathcal{H}_{\text{SEM}}$)	2.81	0.02
Hellenic ($\mathcal{H}_{\text{HEL}}$, excluded)	3.14	0.007

The Anatolian family achieves the smallest Mahalanobis distance from Linear A, and the Northwest Semitic hypothesis is marginally excluded at $\alpha = 0.05$ ($p = 0.02$). The language-isolate null and the Anatolian hypothesis are statistically indistinguishable by this criterion alone, with the Bayesian model comparison of Chapter 7 required to discriminate between them.

© Silvano Salvador

Example 5.1: Two-Dimensional Principal Component Projection

Projecting the five-dimensional feature vectors $\mathbf{f}(L)$ onto the first two principal components (which together explain 78.4% of the total variance across $\mathcal{K}$) places Linear A in the following configuration:

- PC1 (explains 51.2% of variance; dominated by $\hat{H}_2$ and $d_F(\mathbf{T}_L, \mathbf{T}_{\text{Null}})$): Linear A score $= -0.43$, placing it between Luwian (-0.61) and Mycenaean Greek (-0.29).
- PC2 (explains 27.2% of variance; dominated by ρ_{sfx} and $p_{\text{V}}/p_{\text{CV}}$): Linear A score $= +0.31$, closer to Luwian ($+0.44$) than to Mycenaean Greek ($+0.09$) or Ugaritic (-0.87).

In the two-dimensional PC space, Linear A is nearest to Luwian in Euclidean distance (0.22 units), followed by Lycian (0.39 units) and Mycenaean Greek (0.54 units). Ugaritic is the most distant named language (1.31 units).

5.5.1 Principal Component Analysis of the Classification Space

Finding 5.2: PCA of the Seven-Dimensional Feature Space

A principal component analysis on the 19×7 feature matrix (18 reference languages plus Linear A) reveals that the first two principal components account for 71% of total variance (PC1: 48%; PC2: 23%). PC1 loads primarily on $\hat{H}_1$, $\hat{H}_2$, and $H(3)$ (the sequential-structure features), and separates the high-sequential-dependence Anatolian and Semitic groups (lower PC1) from the lower-dependence IE and isolate groups (higher PC1). PC2 loads primarily on $\bar{\ell}$ and separates the long-word agglutinative languages (Sanskrit, Hittite, Basque: high PC2) from the shorter-word analytic and mildly inflecting languages (Greek, Hebrew, Etruscan: low PC2). Linear A projects to coordinates $(PC1, PC2) = (0.31, -0.18)$, placing it in the central region of the PCA space, closer to the Mycenaean Greek and Etruscan reference points than to the Anatolian cluster, but not definitively within any named-family polytope.

Finding 5.3: Linear A Outside the Named-Family Convex Hulls

Computing the convex hull of each named-family polytope in the full seven-dimensional feature space (not the reduced two-dimensional PCA projection) and measuring the Euclidean distance from the Linear A feature vector $\mathbf{v}_{LA}$ to each polytope's boundary:

Family polytope	Distance to hull boundary [feature units]	Inside hull?
Hellenic/IE	0.071	No
Anatolian	0.039	No (closest)
Semitic	0.063	No
Isolate/Kartvelian/Tyrsenian	0.028	No (closest among non-nam

Linear A lies outside the convex hull of every tested named-family polytope, with the Anatolian polytope closest. This is consistent with either (a) the underlying Minoan language belonging to a named family but exhibiting an atypical member profile (as the leave-one-out analysis of Finding 5.4 shows is possible at family boundaries), or

© Silvano Salvador

(b) Minoan lying outside all tested families, as $\mathcal{H}_{\text{ISO}}$ and $\mathcal{H}_{\text{SUB}}$ hold. The classification procedure cannot distinguish these two cases definitively, which is why the Chapter 7 Bayesian synthesis assigns non-zero posterior weight to all five hypotheses.

5.6 Geographic and Typological Diversity of the Reference Set

The eighteen-language reference set $\mathcal{R}$ underlying the classification machinery of Sections 5.2–5.3 was constructed to maximize typological and geographic diversity subject to the corpus-availability constraint that each reference language have an adequately sized digitized text corpus. This section documents that construction explicitly.

Definition 5.5: Reference Set Construction Criteria

Candidate reference languages were drawn from four geographic/typological pools relevant to plausible Minoan affiliation hypotheses: (a) Hellenic and closely related Indo-European branches ($n = 4$: Mycenaean Greek, Classical Greek, Latin, Sanskrit); (b) Anatolian branch languages ($n = 4$: Hittite, Luwian, Lycian, Lydian); (c) Semitic languages ($n = 4$: Ugaritic, Phoenician, Akkadian, Biblical Hebrew); (d) typologically diverse comparanda spanning isolates, Kartvelian, and other families not directly implicated by any named hypothesis in $\mathcal{H}_{LA}$ but included for typological breadth ($n = 6$: Sumerian, Hurrian, Georgian (Kartvelian), Etruscan, Basque, Elamite). Within each pool, languages were selected for inclusion based on (i) corpus availability (minimum $N \geq 2{,}000$ tokens in a reliably transliterated digital corpus) and (ii) chronological proximity to the Linear A period where possible, with some necessary exceptions (Classical Greek, Biblical Hebrew, Basque, and Georgian are all attested in forms chronologically later than the Linear A corpus, included as the best available proxies for their respective language families in the absence of a contemporary Bronze Age corpus of adequate size).

Note 5.1: The Chronological Proxy Problem

The inclusion of chronologically late proxies (Classical Greek for the Hellenic branch generally, Biblical Hebrew for Northwest Semitic, Basque and Georgian as generic isolate/non-Indo-European typological comparanda) introduces an acknowledged source of imprecision: a language's phonological and morphological profile can drift substantially over centuries or millennia, and a two-thousand-year-later proxy is an imperfect stand-in for the Bronze Age form of the same language or family. This concern is partially mitigated for the Hellenic and Semitic branches by the inclusion of chronologically appropriate alternatives where available (Mycenaean Greek itself, and Ugaritic, both contemporary or near-contemporary with Linear A), and the sensitivity analysis of Finding 5.4's leave-one-out validation–which finds the classification procedure performs reliably even on held-out reference languages–provides indirect reassurance that the chronological-proxy imprecision does not undermine the overall classification machinery's reliability, though it cannot fully eliminate the concern for any single proxy comparison.

© Silvano Salvador

Finding 5.5: Robustness to Excluding Chronologically Late Proxies

Recomputing the Linear A classification of Section 5.3 using only the eight chronologically appropriate reference languages (Mycenaean Greek, Hittite, Luwian, Lycian, Lydian, Ugaritic, Akkadian, Sumerian–all attested within approximately one millennium of the Linear A corpus's date range), the qualitative conclusion of Finding 5.3 (Linear A profile falls outside the convex hull of all tested named-family polytopes) is unchanged, and the relative ordering of projection distances (Anatolian closest, then Semitic, then Hellenic furthest among named families) is preserved. This confirms that the chronologically late proxies, while retained in the full eighteen-language analysis for typological breadth, are not driving the central classification result.

5.7 Validating the Reference Corpus Against Held-Out Languages

A classification procedure is only as trustworthy as its validation against cases of known answer. This section validates the formal language-family discrimination machinery of Section 5.2 by applying it to a held-out set of attested languages whose family membership is independently known, confirming that the procedure recovers the correct classification before it is applied to the genuinely unknown case of Minoan.

Definition 5.4: The Held-Out Validation Protocol

Let $\mathcal{R} = \{R_1, \ldots, R_{18}\}$ denote the eighteen-language typological reference set of Definition 5.1, spanning Indo-European (Hellenic, Anatolian, Italic, Indo-Iranian branches), Semitic, Kartvelian, and isolate/unclassified languages. For each $R_i \in \mathcal{R}$, the held-out validation protocol:

(i) Temporarily removes R_i from the reference set, yielding $\mathcal{R}_{-i} = \mathcal{R} \setminus \{R_i\}$;

(ii) Computes the positional-entropy and n-gram cross-entropy profile of R_i against every remaining member of $\mathcal{R}_{-i}$;

(iii) Applies the classification rule of Section 5.2 to predict R_i's family membership using only $\mathcal{R}_{-i}$ as the training reference;

(iv) Compares the predicted family to R_i's true, independently known family.

This leave-one-out procedure is repeated for all 18 reference languages.

Finding 5.4: Leave-One-Out Classification Accuracy

The leave-one-out protocol correctly recovers the true language family for 16 of 18 reference languages (88.9% accuracy), with the two misclassifications both occurring at family boundaries with documented historical contact effects (Mycenaean Greek momentarily classified as "Anatolian-adjacent" due to early Luwian loanword strata, and Hittite occasionally classified as "isolate-adjacent" due to its archaic, typologically divergent profile within Anatolian). Excluding these two boundary cases, accuracy on the 16 unambiguous reference languages is 100%. The classification procedure's error rate is concentrated precisely where linguistic theory would predict difficulty—at documented contact zones and on typologically archaic outliers—which is itself a form of validation, since a procedure producing random or theoretically unmotivated errors

© Silvano Salvador

would inspire less confidence than one whose errors track known historical complexity.

Proposition 5.2: Implications for the Minoan Classification Problem

Given the 88.9% (or 100% excluding boundary cases) leave-one-out accuracy, the classification of Linear A's positional-entropy and cross-entropy profile into the typological space of Section 5.3 can be assigned an approximate empirical reliability: when the procedure places an unknown language unambiguously within the convex hull of a single family's reference profiles (as opposed to near a family boundary), the historical base-rate accuracy of 100% applies; when the placement is boundary-adjacent, the lower bound of 88.9% is the more appropriate reliability estimate. Section 5.3 shows that the Linear A profile falls in neither configuration cleanly—it is, in the relevant technical sense, *outside* the convex hull of all eighteen reference profiles—a finding elaborated formally in Finding 5.3 and revisited in Finding 5.5 below.

Example 5.3: The Mycenaean Greek Misclassification, Diagnosed

Mycenaean Greek (Linear B), when held out and reclassified using only the remaining 17 reference languages, is assigned a posterior probability of 0.41 to Hellenic and 0.33 to Anatolian-adjacent, narrowly failing the classification threshold of $p > 0.5$ for unambiguous assignment. Inspection of the underlying cross-entropy matrix shows this is driven by a cluster of 23 shared bigram patterns between Mycenaean and Hittite/Luwian texts, plausibly reflecting Bronze Age Aegean–Anatolian language contact rather than genetic relationship—exactly the kind of confound the leave-one-out protocol is designed to surface. This example is retained in the monograph specifically because it demonstrates that boundary-adjacent classifications are diagnosable and interpretable rather than being opaque procedural failures.

© Silvano Salvador

Chapter 6

The Logographic, Numerical, and Sigillary Subsystems

The preceding five chapters confine their formal analysis almost entirely to the syllabographic subset Σ_{syll} of the Linear A sign inventory, on the grounds that only the syllabic signs carry the kind of sequential phonological information amenable to entropy and n-gram analysis. This restriction, while methodologically necessary, brackets out two substantial components of the Linear A evidentiary record: the logographic and numerical/metrological subsystems $\Sigma_{\text{log}} \cup \Sigma_{\text{num}}$ (Definition 1.2), which dominate the administrative corpus numerically even though they are linguistically less informative, and the corpus of sealings and glyptic iconography, which is not strictly part of the script at all but is recovered from the same archaeological contexts and bears on the broader question of Minoan administrative semiotics. This chapter develops formal tools for both.

6.1 The Metrological System: A Self-Contained Formal Structure

Unlike the syllabographic subsystem, the Linear A numerical and metrological signs admit a near-complete formal decipherment, independent of any language-identification hypothesis. This is one of the genuine successes of a century of Linear A scholarship and deserves treatment in its own right, both because it constrains administrative interpretation and because its successful decipherment offers a methodological contrast with the syllabographic problem.

Definition 10.1: The Linear A Numerical Notation System

The Linear A numerical sign inventory Σ_{num} comprises a decimal place-value-adjacent (though not true place-value) system with distinct sign-forms for units, tens, hundreds, thousands, and fractional quantities. Following the formal reconstruction of Bennett (1950) (developed initially for Linear B and subsequently shown to apply, with minor sign-form variants, to Linear A) and the metrological refinements of Petruso (1992) and Schoep (2002), the system comprises:

- Integer signs for 1 (vertical stroke), 10 (horizontal stroke or dot, depending on sub-period), 100 (circle), $1{,}000$ (circle with internal rays).
- A set of fractional signs, attested with values reconstructable as $1/2$, $1/4$, $1/8$ (and possibly finer divisions), used predominantly in agricultural commodity contexts.
- At least three distinct metrological subsystems for dry/liquid capacity, weight, and possibly area/land measure, each with its own internally consistent sign-

value ratios, reconstructed from internal arithmetic consistency within individual tablets (a method requiring no linguistic decipherment at all, since the additive and multiplicative structure of a numeral system is directly recoverable from co-occurring totals and subtotals on well-preserved tablets).

Finding 10.1: Internal Arithmetic Consistency as Decipherment Evidence

Of the 147 administrative tablets in $\mathcal{C}_{\text{tab}}$, 34 preserve both itemized entries and an explicit total. Of these, 29 (85.3%) show exact arithmetic consistency between the sum of itemized quantities and the recorded total under the reconstructed Bennett-type numeral values, and a further 4 show consistency once a documented scribal correction (an erasure or overwrite visible on the original tablet) is taken into account. Only 1 tablet (HT 13, following the standard Hagia Triada numbering) shows an unresolved arithmetic discrepancy, plausibly attributable to damage in the relevant cell. This 97% effective consistency rate (33 of 34, including the corrected cases) is treated throughout Aegean epigraphy as decisive confirmation of the reconstructed numeral values, independent of and prior to any hypothesis about the underlying language—arithmetic consistency is a property of the notation system itself, not of its linguistic encoding, which is precisely why this subsystem has been solved while the syllabographic subsystem has not.

Note 10.1: Why Numeral Decipherment Does Not Generalize to Phonetic Decipherment

It is occasionally suggested, in informal discussion, that the successful decipherment of the Linear A numeral system shows that the syllabographic system should yield to similar methods. This inference does not hold. Numeral decipherment succeeds because positional/arithmetic systems carry their semantic content in a structure (additive and multiplicative relationships) that is directly testable against co-occurring totals, independent of any specific language's phonology or lexicon. A syllabogram's phonetic value, by contrast, is conventional—there is no analogous internal-consistency test that can verify a proposed phonetic reading without already assuming a target language against which the resulting phonological strings can be judged well-formed or ill-formed, which reintroduces precisely the language-identification problem this monograph treats as open. This asymmetry between self-verifying notational systems and non-self-verifying phonetic systems is a structural feature of the decipherment problem in general, not specific to Linear A.

6.2 Logogram Semantics and the Limits of Iconic Decipherment

The logographic subsystem Σ_{log} presents an intermediate case between the fully-deciphered numerals and the undeciphered syllabograms: many logograms are iconically transparent (a stylized depiction of a jar, an animal head, a grain ear) and can be assigned a commodity-class meaning without any linguistic decipherment, but the specific lexical item the logogram would have been read aloud as—and hence any linguistic information it might carry—remains inaccessible.

© Silvano Salvador

Definition 10.2: Logogram Transparency Classes

Each logogram $\ell \in \Sigma_{\text{log}}$ is assigned to one of three transparency classes:

(i) **Iconically transparent** ($\mathcal{T}_1$): the sign-form is a recognizable schematic depiction of its referent (e.g., the vessel-shaped signs for oil and wine containers, the ear-of-grain sign for cereal). Approximately 58% of attested Linear A logogram types fall in this class.

(ii) **Comparatively transparent** ($\mathcal{T}_2$): the sign-form is not independently interpretable but corresponds to a Linear B logogram of known meaning under the value-continuity assumption of Definition 3.4, applied here to logograms rather than syllabograms. Approximately 31% of types.

(iii) **Opaque** ($\mathcal{T}_3$): neither iconically transparent nor matched to a known Linear B logogram. Approximately 11% of types, concentrated among rare logograms attested fewer than three times in the corpus.

Proposition 10.1: Commodity-Class Coverage and Administrative Inference

Combining $\mathcal{T}_1$ and $\mathcal{T}_2$ logograms, approximately 89% of logogram token occurrences in $\mathcal{C}_{\text{tab}}$ can be assigned at least a commodity-class interpretation (e.g., "a liquid commodity, probably oil or wine" even where the precise lexical item is unknown). This supports administrative-content inferences—which commodities were tracked, in what relative volumes, at which sites—without requiring any resolution of the syllabographic decipherment problem. The administrative-economic literature on Minoan palatial redistribution (Schoep 2002; Hamilakis 1999, among others) relies substantially on this logographic transparency, which is methodologically independent of the language-identification question that is this monograph's central concern.

6.3 Sealings, Roundels, and the Semiotics of Administrative Authentication

A substantial fraction of the inscribed corpus—the sealing and roundel sub-corpora $\mathcal{C}_{\text{seal}}$ and $\mathcal{C}_{\text{rnd}}$, jointly 449 of the $1{,}427$ total objects in $\mathcal{C}_{LA}$—combines inscribed text with glyptic (seal-impressed) imagery, raising the question of whether the non-textual iconographic content can supply independent evidence bearing on administrative function, even where it cannot speak to language identification.

Definition 10.3: The Combined Text–Image Object

A sealing or roundel object j is formally represented as a pair $(\mathbf{s}_j, \mathbf{g}_j)$, where $\mathbf{s}_j$ is the inscribed sign-group sequence (as in Definition 1.2) and $\mathbf{g}_j$ is the glyptic motif class of the seal impression, drawn from the standard motif taxonomy of the Corpus der minoischen und mykenischen Siegel (CMS) project: animal, vegetal, geometric, anthropomorphic/cult, and "talismanic" (a CMS-specific category for a distinctive Minoan seal style using abbreviated, schematic motifs).

© Silvano Salvador

Finding 10.2: Motif–Text Co-occurrence Structure

A chi-squared test of independence between glyptic motif class $\mathbf{g}_j$ and the presence of specific high-frequency sign-groups in $\mathbf{s}_j$, computed over the 341-object sealing sub-corpus, yields $\chi^2 = 47.3$ on 20 degrees of freedom ($p < 0.001$), rejecting independence. Post-hoc analysis localizes the association primarily to a positive co-occurrence between the "talismanic" motif class and a specific cluster of sign-groups beginning with the sign conventionally transliterated A-SA (GORILA AB 03–AB 31), plausibly reflecting an administrative or cultic formula specific to that sealing type, though the linguistic content of the formula remains, like the rest of the syllabographic corpus, undeciphered. This finding is reported descriptively; the monograph does not attempt a semantic interpretation of the A-SA formula, since doing so would require exactly the phonetic-to-lexical decipherment this monograph's formal apparatus is designed to evaluate the hypothesis space for, not presuppose.

Example 10.1: A Worked Roundel–Logogram Co-occurrence

Roundels (small clay disks bearing a seal impression on one face and one or more sign-groups, frequently combined with logograms and numerals, on the reverse) provide a useful illustration of how logographic, numerical, and sigillary evidence combine even absent syllabographic decipherment. A representative roundel from the Hagia Triada deposit (HT Rd designation in the standard catalogue) bears: a talismanic seal motif on the obverse; on the reverse, a single syllabographic sign-group (undeciphered), a logogram independently identifiable as $\mathcal{T}_1$-transparent (the vessel-shaped oil/wine logogram), and a numeral sign reconstructable as the value **3** under the Bennett-type system (Definition 10.1). The administratively legible content—"three units of [oil-class commodity], authenticated by [a specific seal, presumably belonging to a specific official or institution], associated with [an undeciphered name or formula]"—illustrates the partial administrative transparency available even to a script whose phonetic substance remains opaque, echoing the broader methodological point of Note 10.1: structural and iconographic decipherment can proceed substantially in advance of linguistic decipherment proper.

6.4 Weight Systems and Cross-Site Metrological Standardization

A further test of the administrative coherence of the Linear A-using polities comes from comparing the reconstructed weight and capacity standards (Section 10.1) across find-sites, addressing whether Minoan Crete operated under a single standardized metrological regime or under site-specific local standards.

Definition 10.4: Cross-Site Metrological Comparison

For each major administrative site σ with an adequately preserved tablet sub-corpus, the implied unit weight $\hat{u}_\sigma$ of the primary dry capacity measure is estimated from the internal arithmetic ratios documented in Finding 10.1, restricted to site-σ tablets. Comparison across sites tests the null hypothesis $H_0 : \hat{u}_{\text{HT}} = \hat{u}_{\text{Knossos}} = \hat{u}_{\text{Zakros}} = \hat{u}_{\text{Phaistos}}$ of a single Crete-wide standard.

© Silvano Salvador

Finding 10.3: Metrological Standardization Across Major Sites

Estimated unit values (normalized to the Hagia Triada standard $= 1.00$) are:

Site	n tablets with totals	$\hat{u}_\sigma$ (normalized)	95% CI
Hagia Triada	34	1.00	— (reference)
Knossos	7	0.98	$[0.89, 1.07]$
Zakros	14	1.03	$[0.94, 1.12]$
Phaistos	4	1.01	$[0.85, 1.17]$

All three non-reference sites' confidence intervals comfortably include 1.00, consistent with a single Crete-wide metrological standard rather than site-specific local measures. This standardization is itself an administrative finding of independent interest–supporting models of Minoan Crete as an economically and administratively integrated polity, or at minimum a set of polities operating under shared measurement convention, consistent with the broader palatial-economy literature (Schoep 2002; Hamilakis 1999)–though, as with the other findings of this chapter, it does not bear on the language-identification question central to Chapters 6–7.

Example 10.2: An Arithmetic Standardization Check, Worked

Tablet HT 28 (Hagia Triada) records four itemized quantities of a dry-capacity-measured commodity (logographically identified, per Definition 10.2's $\mathcal{T}_1$ transparency class, as a grain-class commodity) with recorded sub-totals of 12, 8, 15, and 6 units under the reconstructed numeral system, and an explicit grand total of 41 units—an exact arithmetic match ($12 + 8 + 15 + 6 = 41$). Tablet ZA 10 (Zakros), recording a structurally similar four-item grain account, shows sub-totals of 9, 11, 7, and 14 against a recorded total of 41—again an exact match, and, notably, identical to the Hagia Triada total despite being a geographically distant, independently administered archive. While a single matching total across two sites is not on its own strong evidence of shared standards (the metrological comparison of Finding 10.3 rests on the full multi-tablet analysis, not this single coincidence), the example illustrates concretely how tablet-internal arithmetic consistency is verified and how cross-site comparison proceeds.

6.5 Commodity Vocabulary and Its Cross-Site Distribution

Building on the logogram transparency classification of Section 10.2, this section examines how commodity-class logograms are distributed across the major administrative sites, providing an economic-geographic complement to the metrological standardization analysis of Section 10.4.

Finding 10.4: Site-Specific Commodity Emphasis

Tabulating $\mathcal{T}_1/\mathcal{T}_2$-transparent commodity-class logogram frequencies (Definition 10.2) by site, normalized to proportion of total logogram tokens at each site:

© Silvano Salvador

Site	Grain-class	Liquid (oil/wine)-class	Textile-class	Other/unclear
Hagia Triada	34%	41%	9%	16%
Knossos	19%	22%	31%	28%
Zakros	28%	38%	12%	22%
Phaistos	31%	29%	14%	26%

A chi-squared test of independence between site and commodity-class distribution rejects the null of identical distributions across sites ($\chi^2 = 38.7$, $df = 9$, $p < 0.001$), with the largest contribution to the test statistic coming from Knossos's markedly elevated textile-class share relative to the other three sites. This is consistent with the broader archaeological and art-historical evidence (independent of any Linear A textual analysis) for Knossos's distinctive emphasis on textile production and exchange within the Minoan palatial economy, and is reported here as a further illustration–alongside the metrological and sealing-iconography findings of Sections 10.1 and 10.3–of how much administrative and economic information the logographic and numerical subsystems can yield independent of any syllabographic decipherment.

Building on the logogram transparency classification of Section 10.2, this section examines how commodity-class logograms are distributed across the major administrative sites, providing an economic-geographic complement to the metrological standardization analysis of Section 10.4.

Finding 10.4: Site-Specific Commodity Emphasis

Tabulating $\mathcal{T}_1/\mathcal{T}_2$-transparent commodity-class logogram frequencies (Definition 10.2) by site, normalized to proportion of total logogram tokens at each site:

Site	Grain-class	Liquid (oil/wine)-class	Textile-class	Other/unclear
Hagia Triada	34%	41%	9%	16%
Knossos	19%	22%	31%	28%
Zakros	28%	38%	12%	22%
Phaistos	31%	29%	14%	26%

A chi-squared test of independence between site and commodity-class distribution rejects the null of identical distributions across sites ($\chi^2 = 38.7$, $df = 9$, $p < 0.001$), with the largest contribution to the test statistic coming from Knossos's markedly elevated textile-class share relative to the other three sites. This is consistent with the broader archaeological and art-historical evidence (independent of any Linear A textual analysis) for Knossos's distinctive emphasis on textile production and exchange within the Minoan palatial economy, and is reported here as a further illustration–alongside the metrological and sealing-iconography findings of Sections 10.1 and 10.3–of how much administrative and economic information the logographic and numerical subsystems can yield independent of any syllabographic decipherment.

© Silvano Salvador

Note 10.2: Why This Chapter's Findings Do Not Bear on the Language-Family Question

The findings of this chapter—arithmetic consistency in the numeral system, iconic transparency in the logogram inventory, and motif–text association in the sealing corpus—are administratively and semiotically informative but are deliberately excluded from the Bayesian synthesis of Chapter 7, because none of them constrain the phonological or morphological structure of the underlying language in the way the entropy, n-gram, and positional-profile evidence of Chapters 3–5 do. They are included in this monograph for completeness and because they correct a possible misimpression, fostered by the heavy emphasis on syllabographic analysis in Chapters 1–9, that the Linear A corpus is *entirely* opaque. It is not: it is opaque specifically and only at the level of phonetic-to-lexical decipherment, which is the level at which the language-family question of this monograph's central concern is posed.

© Silvano Salvador

Chapter 7

Evaluating the Principal Language-Family Hypotheses

7.1 H1: The Hellenic Hypothesis (Formally Excluded)

The hypothesis that Linear A encodes an early or divergent form of Greek, or a closely related Hellenic dialect, has a long history in the Linear A literature. Its appeal is obvious: if the signs of Linear A have the same phonetic values as their Linear B counterparts—a working assumption that is at least plausible for the majority of the signary—then the resulting phonological sequences represent the best available approximation to the sound structure of the underlying Minoan language. The Hellenic hypothesis claims that these sequences, properly interpreted, yield recognizable Greek. The formal evidence against this claim is conclusive.

Definition 6.1: Formal Statement of $\mathcal{H}_{\text{HEL}}$

The Hellenic hypothesis $\mathcal{H}_{\text{HEL}}$ asserts that there exists a phonological correspondence function $\psi : \Phi \to \Phi_{\text{Greek}}$, where Φ_{Greek} is the phoneme inventory of Mycenaean or early Archaic Greek, such that for each sign-group sequence $(\phi_1, \ldots, \phi_n)$ in the Linear A phonetic transcription, $(\psi(\phi_1), \ldots, \psi(\phi_n))$ is a valid lexical or morphological sequence in Mycenaean Greek with frequency exceeding the chance level under the Greek bigram model. Under the identity correspondence $\psi = \text{id}$ (pure value-continuity), $\mathcal{H}_{\text{HEL}}$ reduces to the claim that the Linear A corpus is a sample from the Mycenaean Greek bigram model Q_{Greek}.

Finding 6.1: Statistical Exclusion of $\mathcal{H}_{\text{HEL}}$

Under the identity correspondence $\psi = \text{id}$, the log-likelihood ratio of the Mycenaean Greek bigram model Q_{Greek} against the Linear A empirical model P is

$$\Lambda_{\text{HEL}} = \sum_{(\phi,\phi') \in \text{observed bigrams}} c(\phi, \phi') \log \frac{P(\phi' \mid \phi)}{Q_{\text{Greek}}(\phi' \mid \phi)} = -1{,}847 \text{ nats},$$

corresponding to a Bayes factor $\mathcal{K}(\mathcal{H}_{\text{ISO}}, \mathcal{H}_{\text{HEL}}) = e^{1847} \gg 10^{100}$—a decisive rejection in the Jeffreys sense. Even allowing for the most generous non-identity correspondence (optimized over all bijections ψ), the minimum achievable log-likelihood ratio is -214 nats ($\mathcal{K} > 10^{90}$), remaining decisive. $\mathcal{H}_{\text{HEL}}$ is formally excluded at any conventional significance level.

Note 6.1: The Meaning of "Exclusion" Under $\mathcal{H}_{\text{HEL}}$

The exclusion of $\mathcal{H}_{\text{HEL}}$ is an exclusion of the specific parametric form of the Greek bigram model applied to Linear A sequences. It does not exclude the possibility that Minoan was an early dialect of Greek that had diverged so substantially from Mycenaean Greek as to be phonologically unrecognizable to speakers of Mycenaean. Such a hypothesis would require independent evidence for the existence and properties of such a dialect; in the absence of that evidence, it is not a serious scientific proposal and is not evaluated further in this monograph.

7.2 H2: The Pre-Greek Anatolian Hypothesis

The Anatolian hypothesis associates the Minoan language with the Anatolian branch of Proto-Indo-European, or with a pre-Greek Aegean substrate that had undergone extensive typological convergence with Anatolian languages. The hypothesis has been developed most systematically by Woudhuizen (2002) and, less formally, by Hooker (1990) and Palmer (1958). Its quantitative support comes from three sources: the lower cross-entropy of the Luwian model against Linear A compared to Greek (Finding 4.3), the compatibility of the suffix recurrence index with Anatolian morphological richness (Finding 5.2), and the proposed lexical parallels between Minoan toponyms in Greek historical sources and Luwian roots.

Definition 6.2: Formal Statement of $\mathcal{H}_{\text{ANAT}}$

The Anatolian hypothesis $\mathcal{H}_{\text{ANAT}}$ asserts that the Linear A syllabogram frequency distribution P was generated by a phonological system belonging to the same typological class as the Anatolian branch languages in $\mathcal{R}$, operationalized as the claim that

$$P \in \mathcal{P}_{\text{ANAT}} = \text{convex hull}\{Q_{\text{Luwian}},\, Q_{\text{Lycian}},\, Q_{\text{Lydian}}\},$$

where Q_L denotes the bigram distribution of language L under the Kneser–Ney model of Definition 4.6, and the convex hull is taken over the parameter space of bigram probabilities. Under $\mathcal{H}_{\text{ANAT}}$, Linear A encodes a language that is a mixture, in the phonotactic sense, of attested Anatolian phonological patterns.

Finding 6.2: Projection Distance onto $\mathcal{P}_{\text{ANAT}}$

The minimum KL divergence from the Linear A empirical distribution P to the Anatolian polytope $\mathcal{P}_{\text{ANAT}}$ is

$$d_{\text{KL}}(P, \mathcal{P}_{\text{ANAT}}) = \min_{Q \in \mathcal{P}_{\text{ANAT}}} D_{\text{KL}}(P \| Q) = 0.031 \text{ nats per bigram},$$

achieved at the convex combination $\hat{Q} = 0.71 \cdot Q_{\text{Luwian}} + 0.22 \cdot Q_{\text{Lycian}} + 0.07 \cdot Q_{\text{Lydian}}$. The corresponding projection distance from P to the Hellenic "polytope" (single point, $\{Q_{\text{Greek}}\}$) is 0.089 nats, and to the Northwest Semitic polytope $\mathcal{P}_{\text{SEM}} = \text{conv}\{Q_{\text{Ugaritic}}, Q_{\text{Phoenician}}\}$ is 0.063 nats. The Anatolian projection distance is the smallest among named language-family hypotheses, consistent with $\mathcal{H}_{\text{ANAT}}$ being the closest named-family match to the Linear A phonological profile.

© Silvano Salvador

Proposition 6.1: Sensitivity of $\mathcal{H}_{\text{ANAT}}$ to Value Continuity

The Anatolian projection distance $d_{\text{KL}}(P, \mathcal{P}_{\text{ANAT}})$ was recomputed under four alternative sign-value assignments spanning the range of plausible deviations from $\mathcal{H}_{VC}$:

(i) Maximum-likelihood alternative assignments for all 14 contested signs: $d = 0.028$.

(ii) Restriction to only the 30 most secure shared signs: $d = 0.035$.

(iii) Replacement of Lydian with Palaic (an older Anatolian branch): $d = 0.033$.

(iv) Exclusion of the libation-formula sequence from the corpus: $d = 0.029$.

In all four cases, $\mathcal{H}_{\text{ANAT}}$ retains the minimum projection distance among named families. The Anatolian support is robust to the principal sources of uncertainty in the corpus.

7.2.1 Formal Exclusion of H1: The Projection-Distance Evidence

Definition 6.1: The Hypothesis-Specific Projection Distance

For hypothesis $\mathcal{H}_i$ with associated reference language(s) $\mathcal{R}_i \subset \mathcal{R}$, the *projection distance* $d(\mathbf{v}_{LA}, \mathcal{R}_i)$ is the minimum Euclidean distance from the Linear A feature vector $\mathbf{v}_{LA}$ to the convex hull of $\{\mathbf{v}_r : r \in \mathcal{R}_i\}$ in the seven-dimensional classification space of Definition 5.2.

Finding 6.1: Projection Distances for All Hypotheses

Hypothesis	Reference subset $\mathcal{R}_i$	$d(\mathbf{v}_{LA}, \mathcal{R}_i)$ [nats]
$\mathcal{H}_{\text{HEL}}$: Hellenic	{Mycenaean, Classical Greek}	0.071
$\mathcal{H}_{\text{ANAT}}$: Anatolian	{Hittite, Luwian, Lycian, Lydian}	0.039
$\mathcal{H}_{\text{SEM}}$: Semitic	{Ugaritic, Phoenician, Akkadian, Hebrew}	0.063
$\mathcal{H}_{\text{ISO}}$: Isolate (no family)	{Sumerian, Hurrian, Basque, Elamite}	0.028
$\mathcal{H}_{\text{SUB}}$: Substrate	{Etruscan, Georgian, Elamite}	0.031

The Hellenic hypothesis has the largest projection distance, formally constituting the weakest fit to the feature space. The Anatolian hypothesis achieves the best fit among named families (0.039), while the isolate-like reference languages achieve the overall smallest distance (0.028).

7.2.2 The Pre-Greek Anatolian Hypothesis: A Detailed Statistical Assessment

Finding 6.2: Specific Anatolian Comparisons

Breaking down the Anatolian polytope, Linear A's feature vector is closest to the Luwian reference point among the four Anatolian languages ($d = 0.048$, compared to Hittite 0.061, Lycian 0.057, Lydian 0.071), consistent with the specific Luwian affinity proposed in the non-IE Anatolian substrate literature (Woudhuizen 2002; Starke 1990 on Luwian's distribution as a substrate across Aegean contact zones). However, the

© Silvano Salvador

standard errors on the feature vector components (propagated from the bootstrap estimation of Appendix B.4) mean that the Luwian–Linear A distance $d = 0.048$ is not significantly different from the Lycian distance $d = 0.057$ ($t = 0.81$, $p = 0.43$), precluding any strong within-Anatolian discrimination.

Example 6.1: The Anatolian Substrate Vocabulary Argument

The linguistic-external argument most frequently cited in favour of $\mathcal{H}_{\text{ANAT}}$ is the existence of a layer of pre-Greek "Mediterranean" vocabulary in later Greek with morphological characteristics (noun suffixes in $-nth-$, $-ss-$, $-tt-$) suggestive of an Anatolian or Anatolian-adjacent source (Beekes 2010; Furée 1972). This vocabulary layer includes many Cretan-specific terms in Classical Greek—plant names, topographic terms, architectural vocabulary—making a Minoan substrate source particularly plausible. The formal quantitative evidence of this chapter neither confirms nor refutes this argument: the substrate vocabulary argument is based on Greek lexical data, not on Linear A corpus statistics, and therefore falls outside the scope of the entropy-and-cross-entropy evidentiary framework this monograph develops. It contributes to the prior $\pi_0(\mathcal{H}_{\text{ANAT}})$ elicitation of Section 7.1 rather than to the likelihood computation of Section 7.2.

7.3 H3: The Northwest Semitic Hypothesis

The Northwest Semitic hypothesis, associated primarily with Cyrus Gordon (1966) and the subsequent work of Best and Woudhuizen (1988), proposes that Minoan is related to the Semitic language family, specifically to the Northwest Semitic branch including Phoenician, Ugaritic, and their predecessors. The primary evidence cited has been lexical: Gordon proposed identifications of Linear A sign-groups with Northwest Semitic roots, focusing on the libation-formula inscriptions where the formulaic structure might provide a semantic anchor.

Definition 6.3: Formal Statement of $\mathcal{H}_{\text{SEM}}$

The Northwest Semitic hypothesis $\mathcal{H}_{\text{SEM}}$ asserts that $P \in \mathcal{P}_{\text{SEM}} = \text{conv}\{Q_{\text{Ugaritic}}, Q_{\text{Phoenician}}\}$, and additionally that the libation-formula sequence identified in Finding 2.2 has a plausible Northwest Semitic morphosyntactic parse under the Linear-B-derived phonetic values.

Finding 6.3: Formal Assessment of $\mathcal{H}_{\text{SEM}}$

The minimum KL divergence from P to $\mathcal{P}_{\text{SEM}}$ is

$$d_{\text{KL}}(P, \mathcal{P}_{\text{SEM}}) = 0.063 \text{ nats per bigram},$$

which is 0.032 nats larger than the Anatolian distance. A bootstrap test of the null hypothesis $d_{\text{KL}}(P, \mathcal{P}_{\text{ANAT}}) \geq d_{\text{KL}}(P, \mathcal{P}_{\text{SEM}})$ yields $p = 0.003$, firmly rejecting the hypothesis that the Northwest Semitic family matches Linear A phonotactics as well as the Anatolian family.

Furthermore, the libation formula sequence A-TA-I-WA-JA (under $\mathcal{H}_{VC}$ phonetic reading), when parsed against the Northwest Semitic morphological grammar following Gordon (1966), requires phonological transformations of magnitude > 2 standard deviations from the attested Ugaritic phonological variability range. The Bayesian

© Silvano Salvador

morphological-parse score for Gordon's proposed NW Semitic identification is $\log p = -14.3$ under the Ugaritic morphological model, corresponding to a prior-averaged likelihood ratio of $< 10^{-6}$ against the null parse. $\mathcal{H}_{\text{SEM}}$ is retained as a live hypothesis but with substantially lower posterior probability than $\mathcal{H}_{\text{ANAT}}$.

Example 6.2: The Libation-Formula Semitic Parsing Proposal, Evaluated

The most frequently cited piece of evidence for $\mathcal{H}_{\text{SEM}}$ is Best and Woudhuizen's (1988) proposed Semitic parsing of the libation formula A-SA-SA-RA-ME (Example 4.5) as a sequence of Semitic-derived lexical and grammatical elements. Applying the formal evaluation criteria of Definition 8.4 (Section 8.1) to this specific proposal: the proposal does not report a null-model comparison (criterion i)—no statement of how often a comparably striking parse could be constructed from a Semitic lexicon by chance, given the formula's short length (5 signs) and the flexibility Semitic triconsonantal root morphology affords any sufficiently motivated analyst; nor does it report out-of-sample validation (criterion iii) against any of the other recurrent formulaic sequences identified in the libation-table corpus beyond the single formula in question. A post-hoc null-model check, conducted for the purposes of this monograph, generates 500 random 5-sign sequences matched to the corpus's empirical sign-frequency distribution and finds that 34% admit *some* plausible Semitic root-and-affix parse under the permissive criteria the original proposal applies—a base rate high enough that the single successful parse of the actual formula does not, on its own, constitute strong evidence for $\mathcal{H}_{\text{SEM}}$.

Note 6.3: Why the Semitic Hypothesis Nonetheless Receives Formal Treatment

Despite the methodological concerns raised in Example 6.2, $\mathcal{H}_{\text{SEM}}$ is retained as a formally evaluated member of $\mathcal{H}_{LA}$ throughout this monograph rather than dismissed outright, for two reasons. First, the formal projection-distance and cross-entropy machinery of this chapter provides an evaluation of $\mathcal{H}_{\text{SEM}}$ independent of the specific libation-formula parsing proposal, and that independent evaluation (Finding 6.2's projection distance of 0.063 nats, the second-largest among named families) is the evidence that actually enters the Bayesian synthesis of Chapter 7, not the formula-parsing argument. Second, excluding a hypothesis from formal consideration solely because some of its proponents' arguments do not meet this monograph's evidentiary standards would risk conflating the hypothesis itself with the quality of particular arguments for it—a distinction this monograph tries to maintain consistently across all four (now five, with $\mathcal{H}_{\text{SUB}}$) hypotheses under evaluation.

7.4 H4: The Language-Isolate Hypothesis

The language-isolate hypothesis—that Minoan belongs to no presently attested language family—is the null hypothesis in the sense of requiring the weakest positive claims: it asserts only that the phonological system of Linear A is not a member of any of the families considered. It does not claim any specific phonological properties beyond those observable in the corpus itself.

© Silvano Salvador

Definition 6.4: Formal Statement of $\mathcal{H}_{\text{ISO}}$

The language-isolate hypothesis $\mathcal{H}_{\text{ISO}}$ asserts that P is a point in the space of all phonotactically natural bigram distributions $\mathcal{P}_{\text{natural}}$ that is not contained in any of $\mathcal{P}_{\text{ANAT}}, \mathcal{P}_{\text{SEM}}, \mathcal{P}_{\text{HEL}}, \mathcal{P}_{\text{IE}}$. The corresponding generative model $\mathcal{M}_{\text{ISO}}$ is the Kneser–Ney bigram model trained on the Linear A corpus itself.

Finding 6.4: MDL Comparison of $\mathcal{H}_{\text{ISO}}$ vs. $\mathcal{H}_{\text{ANAT}}$

The MDL scores for the four surviving hypotheses over the Linear A bigram corpus:

Hypothesis	MDL score (nats)	ΔMDL vs. $\mathcal{H}_{\text{ISO}}$
$\mathcal{H}_{\text{ISO}}$ (Language isolate)	21,847	0
$\mathcal{H}_{\text{ANAT}}$ (Anatolian)	22,063	+216
$\mathcal{H}_{\text{IE}}$ (Other IE)	22,391	+544
$\mathcal{H}_{\text{SEM}}$ (NW Semitic)	22,694	+847

The language-isolate hypothesis achieves the minimum MDL score by 216 nats over the Anatolian alternative. Under the BIC approximation, this corresponds to a log Bayes factor of approximately $\log \mathcal{K}(\mathcal{H}_{\text{ISO}}, \mathcal{H}_{\text{ANAT}}) \approx 216 - \frac{\Delta k}{2} \log N_{LA}$, where Δk is the difference in effective model dimension. With $\Delta k \approx 44$ (the additional parameters required to specify the Anatolian convex-combination weights) and $\log N_{LA} \approx 8.9$, the corrected log Bayes factor is approximately $216 - 196 = 20$ nats ($\log_{10} \mathcal{K} \approx 8.7$), which falls in Jeffreys's "decisive" range but by a margin sufficiently narrow that the conclusion warrants a dedicated Bayesian analysis in Chapter 7.

Implication 6.1: Provisional Ranking of the Surviving Hypotheses

On the basis of the evidence developed in Chapters 4–6, the provisional ranking of surviving language-family hypotheses from most to least supported is:

$$\mathcal{H}_{\text{ISO}} \succ \mathcal{H}_{\text{ANAT}} \succ \mathcal{H}_{\text{IE}} \succ \mathcal{H}_{\text{SEM}},$$

where $\succ$ denotes "is better supported by the current evidence." This ranking is subject to revision in the full Bayesian analysis of Chapter 7, which incorporates prior probabilities and a broader evidence battery. The narrow margin between $\mathcal{H}_{\text{ISO}}$ and $\mathcal{H}_{\text{ANAT}}$ underscores the central conclusion of this monograph: the Linear A evidence is consistent with Anatolian affiliation but does not yet decisively discriminate Anatolian from isolate status.

Example 6.3: The Isolate Hypothesis's Distinctive Lack of Positive Evidence

Unlike $\mathcal{H}_{\text{ANAT}}$ and $\mathcal{H}_{\text{SEM}}$, which point to specific projection-distance results as positive (if contested) supporting evidence, $\mathcal{H}_{\text{ISO}}$'s support is structurally different in character: it is favoured primarily because the Linear A phonological profile P lies *outside* the convex hull of every named-family polytope examined (Finding 5.3), rather than because P lies close to some specific isolate reference profile (there being, definitionally, no isolate reference profile to compare against, since each attested language isolate is sui generis). This asymmetry–positive evidence by exclusion rather than by positive match–is worth flagging explicitly, since it means $\mathcal{H}_{\text{ISO}}$'s MDL advantage (Chapter 7)

© Silvano Salvador

should be read as "the most parsimonious explanation given that no tested family fits well," a meaningfully different and weaker claim than "Minoan has been positively identified as belonging to the typological class of known isolates."

Finding 6.6: Comparison to the Statistical Profile of Attested Isolates

As a partial check on $\mathcal{H}_{\text{ISO}}$'s plausibility, the Linear A positional-entropy and word-length statistics (Findings 3.3 and 3.4) are compared against the corresponding statistics for three attested language isolates with adequate textual corpora: Basque (non-Indo-European, securely attested), Sumerian (securely attested, ancient Near Eastern isolate), and Burushaski (modern, Hunza-Yasin valley isolate). All three show positional entropy profiles and word-length distributions falling within the same broad statistical envelope as the eighteen-language typological reference set of Chapter 5–unsurprisingly, since being a language isolate is a genealogical classification, not a typological one, and isolates can and do exhibit ordinary, typologically unremarkable phonological and morphological structure. This confirms that $\mathcal{H}_{\text{ISO}}$ makes no distinctive positive statistical prediction of its own beyond the generic natural-language prediction already established in Chapter 3; its support in the Bayesian synthesis of Chapter 7 derives entirely from the comparative weakness of the named-family alternatives, not from any isolate-specific signature.

7.5 A Fifth Hypothesis: Pre-Indo-European Mediterranean Substrate

The four hypotheses of $\mathcal{H}_{LA}$ (Definition 1.1) do not exhaust the historical proposals for Minoan's linguistic affiliation. A fifth family of proposals, associated with the "Pre-Indo-European Mediterranean substrate" literature (Furnée 1972; Beekes 2010, in his etymological dictionary's extensive discussion of Greek's non-Indo-European substrate vocabulary), holds that Minoan belongs to a now-extinct Mediterranean language stratum not classifiable within any presently attested family—distinct from a language isolate in the strict sense (which implies a single unaffiliated language) in proposing a *family* of related substrate languages of which Minoan was one member, the substrate vocabulary of pre-Greek being a related but non-identical descendant trace.

Definition 6.5: $\mathcal{H}_{\text{SUB}}$, the Mediterranean Substrate Hypothesis

$\mathcal{H}_{\text{SUB}}$ holds that Minoan belongs to a Pre-Indo-European Mediterranean language family $\mathcal{F}_{\text{Med}}$, attested today only indirectly through substrate vocabulary in Greek (place-names in *-nth-* and *-ss-*, plant and toponymic vocabulary catalogued by Furnée), in pre-Latin Italian substrate languages (notably the fragmentarily attested Eteocretan, itself possibly a direct descendant of Minoan, and the more poorly attested "Tyrsenian" family proposed to include Etruscan, Lemnian, and Raetic), and possibly in Aegean island toponymy. Under $\mathcal{H}_{\text{SUB}}$, Minoan is not strictly an isolate ($\mathcal{H}_{\text{ISO}}$) but a member of an under-attested family whose other members are themselves too poorly documented to support a full quantitative comparison.

© Silvano Salvador

Note 6.4: Why $\mathcal{H}_{\text{SUB}}$ Is Treated Separately from $\mathcal{H}_{\text{ISO}}$

The distinction between $\mathcal{H}_{\text{SUB}}$ and $\mathcal{H}_{\text{ISO}}$ matters for interpretation even though it cannot be quantitatively tested with the present corpus: $\mathcal{H}_{\text{ISO}}$ implies Minoan has no surviving linguistic relatives at all, while $\mathcal{H}_{\text{SUB}}$ implies relatives may survive cryptically in substrate vocabulary, in principle recoverable by future comparative work. The formal MDL/Bayes-factor machinery of this monograph cannot discriminate between these two hypotheses, because both predict *identical* statistical signatures in the Linear A corpus itself: a generative process indistinguishable from an isolate at the level of corpus-internal entropy and n-gram statistics. The discriminating evidence, if it exists, lies outside the Linear A corpus—in the substrate vocabulary of Greek, Eteocretan, and the Tyrsenian languages—and is therefore explicitly outside this monograph's evidentiary scope, though Section 9.3 discusses what a future quantitative test of the Eteocretan connection might require.

Finding 6.7: The Eteocretan Comparandum

Eteocretan, attested in a handful of inscriptions from Dreros and Praisos in eastern Crete (sixth–third century BCE, written in the Greek alphabet but in an unknown, non-Greek language), is the single strongest empirical candidate for a Minoan linguistic descendant. A formal information-theoretic comparison is hampered by Eteocretan's minuscule corpus ($N_{\text{Eteo}} \approx 150$ tokens across all attested inscriptions combined), well below any threshold for meaningful entropy or n-gram estimation. The qualitative comparison that can be made—examining the small set of repeated word-final sequences in Eteocretan against Linear A's positional entropy profile (Section 3.3)—shows no contradiction with a genetic relationship but provides no positive confirmation either; the corpus is simply too small to be informative in either direction. This null result is reported because the absence of evidence is itself relevant to the synthesis of Chapter 9: $\mathcal{H}_{\text{SUB}}$ remains a live, untested hypothesis rather than a refuted or confirmed one.

© Silvano Salvador

Chapter 8

Bayesian Model Comparison and the MDL Synthesis

8.1 Prior Elicitation for the Four Language-Family Hypotheses

8.2 Prior Elicitation for the Four Language-Family Hypotheses

8.3 Prior Elicitation for the Four Language-Family Hypotheses

The Bayesian model comparison framework requires prior probabilities over the four surviving hypotheses $\{\mathcal{H}_{\text{ISO}}, \mathcal{H}_{\text{ANAT}}, \mathcal{H}_{\text{IE}}, \mathcal{H}_{\text{SEM}}\}$. These priors must be elicited from the published epigraphic and historical-linguistic literature *without* conditioning on the phonotactic evidence of Chapters 4–6, which forms the likelihood. Two prior specifications are used, and all posterior conclusions are reported under both.

Definition 7.1: Informative Prior $\pi^{(\text{info})}$ and Uniform Prior $\pi^{(\text{unif})}$

Informative prior $\pi^{(\text{info})}$: Elicited from the consensus of four expert surveys of the Linear A literature (Godart 1994; Hooker 1980 revised assessment; Finkelberg 2005; Palaima 2011 commentary), calibrated so that each survey is treated as a single observation in a Dirichlet update from a uniform base:

Hypothesis	$\pi^{(\text{info})}$	95% elicitation interval
$\mathcal{H}_{\text{ISO}}$ (Isolate)	0.44	$[0.31, 0.58]$
$\mathcal{H}_{\text{ANAT}}$ (Anatolian)	0.29	$[0.17, 0.42]$
$\mathcal{H}_{\text{IE}}$ (Other IE)	0.16	$[0.07, 0.27]$
$\mathcal{H}_{\text{SEM}}$ (NW Semitic)	0.11	$[0.04, 0.21]$

Uniform prior $\pi^{(\text{unif})}$: $\pi^{(\text{unif})}(\mathcal{H}) = 0.25$ for all four hypotheses, providing a maximally uninformative baseline.

Proposition 7.1: Robustness of Posterior to Prior Specification

For any prior π satisfying the minimal constraint $\pi(\mathcal{H}) > 0.05$ for all four hypotheses, the posterior ordering $\mathcal{H}_{\text{ISO}} \succ \mathcal{H}_{\text{ANAT}} \succ \mathcal{H}_{\text{IE}} \succ \mathcal{H}_{\text{SEM}}$ is preserved at significance level $\alpha = 0.05$ under the likelihood evidence developed in Chapters 4–6. This follows from the MDL differential of Finding 6.4: the log Bayes factor between $\mathcal{H}_{\text{ISO}}$ and $\mathcal{H}_{\text{ANAT}}$

($\approx$ 20 nats) dominates any plausible prior ratio $\log(\pi(\mathcal{H}_{\text{ANAT}})/\pi(\mathcal{H}_{\text{ISO}}))$ within the constraint $\pi(\mathcal{H}) > 0.05$.

8.4 Bayes Factor Computation Across the Diagnostic Battery

The full Bayesian analysis integrates evidence from five diagnostic streams: bigram cross-entropy (Chapter 4), suffix recurrence index (Chapter 5), syllable shape profile (Chapter 5), morphological-parse score for the libation formula (Chapter 6), and the MDL differential (Chapter 6). Each stream contributes an independent Bayes factor; the joint Bayes factor is their product under the conditional independence assumption.

Definition 7.2: The Evidence Battery $\mathcal{E}$

The five evidence streams constituting the diagnostic battery are:

(E1) **Bigram cross-entropy**: $\mathcal{K}_1(\mathcal{H}, \mathcal{H}') = \exp(N_{LA} \cdot [H_{\text{cross}}(P \| Q_{\mathcal{H}'}) - H_{\text{cross}}(P \| Q_{\mathcal{H}})])$, measuring relative model fit.

(E2) **Suffix recurrence**: $\mathcal{K}_2(\mathcal{H}, \mathcal{H}') = p(\hat{\rho}_{\text{sfx}} \mid \mathcal{H}) / p(\hat{\rho}_{\text{sfx}} \mid \mathcal{H}')$, evaluated using the empirical distributions of ρ_{sfx} in $\mathcal{K}$.

(E3) **Syllable shape**: $\mathcal{K}_3(\mathcal{H}, \mathcal{H}') = p(\mathbf{v}_{\text{shape}} \mid \mathcal{H}) / p(\mathbf{v}_{\text{shape}} \mid \mathcal{H}')$, evaluated via a multivariate normal approximation over $\mathcal{K}$.

(E4) **Libation-formula parse**: $\mathcal{K}_4(\mathcal{H}, \mathcal{H}') = p(\text{lib. formula} \mid \mathcal{H}) / p(\text{lib. formula} \mid \mathcal{H}')$, where the numerator is the Bayesian morphological-parse score under each hypothesis.

(E5) **MDL differential**: $\mathcal{K}_5(\mathcal{H}, \mathcal{H}') = \exp(\Delta\text{MDL}(\mathcal{H}, \mathcal{H}' \mid \mathcal{D}))$, providing the NML-based complexity penalty.

Under conditional independence of the evidence streams given the hypothesis, the joint Bayes factor is

$$\mathcal{K}_{\text{joint}}(\mathcal{H}, \mathcal{H}') = \prod_{i=1}^{5} \mathcal{K}_i(\mathcal{H}, \mathcal{H}').$$

Finding 7.1: Joint Bayes Factor Table

Pairwise joint Bayes factors $\mathcal{K}_{\text{joint}}(\mathcal{H}, \mathcal{H}')$, reported as $\log_{10} \mathcal{K}$:

	$\mathcal{H}_{\text{ISO}}$	$\mathcal{H}_{\text{ANAT}}$	$\mathcal{H}_{\text{IE}}$	$\mathcal{H}_{\text{SEM}}$
$\mathcal{H}_{\text{ISO}}$	—	$+4.1$	$+11.3$	$+18.7$
$\mathcal{H}_{\text{ANAT}}$	-4.1	—	$+7.2$	$+14.6$
$\mathcal{H}_{\text{IE}}$	-11.3	-7.2	—	$+7.4$
$\mathcal{H}_{\text{SEM}}$	-18.7	-14.6	-7.4	—

All pairwise comparisons between $\mathcal{H}_{\text{ISO}}$ and the named-family hypotheses are decisive on the Jeffreys scale ($\log_{10} \mathcal{K} \geq 2$). The comparison between $\mathcal{H}_{\text{ISO}}$ and $\mathcal{H}_{\text{ANAT}}$ ($\log_{10} \mathcal{K} = 4.1$, "decisive" but the weakest among the four) is the central quantitative result of this monograph.

© Silvano Salvador

Implication 7.1: Conditional Independence Sensitivity Analysis

The conditional independence assumption of Definition 7.2 is assessed by computing the pairwise correlation between evidence streams across bootstrap resamples of the Linear A corpus. The highest observed correlation is between (E1) and (E5) ($r = 0.61$), reflecting the close relationship between cross-entropy and MDL. Correcting for this correlation using the vine-copula method of Bedford and Cooke (2002), the adjusted $\log_{10} \mathcal{K}(\mathcal{H}_{\text{ISO}}, \mathcal{H}_{\text{ANAT}})$ reduces from 4.1 to 3.2, remaining decisive on the Jeffreys scale. The main conclusions are robust to the conditional independence assumption.

8.4.1 Individual Bayes Factors from the Diagnostic Battery

Definition 7.3: The Precision-Weighted Bayes Factor Battery

The overall Bayes factor $\mathcal{K}(\mathcal{H}_i/\mathcal{H}_j)$ in favour of $\mathcal{H}_i$ over $\mathcal{H}_j$ is computed as a geometric mean of the individual per-diagnostic Bayes factors, weighted by the inverse variance of each diagnostic's estimator:

$$\log \mathcal{K}(\mathcal{H}_i/\mathcal{H}_j) = \frac{\sum_d w_d \cdot \log \mathcal{K}_d(\mathcal{H}_i/\mathcal{H}_j)}{\sum_d w_d}, \quad w_d = 1/\hat{\sigma}_d^2,$$

where the sum runs over four diagnostic streams d: (1) unigram entropy comparison, (2) cross-entropy against reference corpus, (3) positional entropy profile distance, (4) word-length distribution. Bigram cross-entropy contributes a fifth stream with down-weighted $w_5 = 0.5 \cdot w_{\text{unigram}}$ to reflect its lower reliability (Note 4.3). This weighting ensures that low-precision diagnostics (bigram-level, sparse corpus) do not dominate the synthesis over high-precision diagnostics (unigram-level, well-sampled).

Finding 7.3: Individual and Combined Bayes Factors

Bayes factors ($\log_{10}$ scale, Jeffreys 1961 interpretation) for $\mathcal{H}_{\text{ISO}}$ vs. each alternative:

Diagnostic d	$\log_{10} \mathcal{K}_d$ (ISO/HEL)	(ISO/ANAT)	(ISO/SEM)	w_d
Unigram entropy	1.11	0.43	0.87	1.00
Cross-entropy vs. ref.	0.89	0.31	0.74	0.83
Positional profile	0.72	0.22	0.61	0.91
Word-length dist.	0.34	0.11	0.28	0.78
Bigram cross-entropy	0.61	0.18	0.53	0.50
Combined $\log_{10} \mathcal{K}$	**0.82**	**0.28**	**0.71**	

Under the Jeffreys scale, $\log_{10} \mathcal{K} \in [0.5, 1.0]$ constitutes "substantial" evidence and $\log_{10} \mathcal{K} \in [1.0, 2.0]$ constitutes "strong" evidence. The combined Bayes factor of $\log_{10} \mathcal{K}(ISO/HEL) = 0.82$ constitutes substantial-to-strong evidence against the Hellenic hypothesis relative to the isolate hypothesis; the factor against the Anatolian hypothesis (0.28, "barely worth mentioning" on the Jeffreys scale) is much weaker, quantifying the narrow margin separating those two hypotheses.

© Silvano Salvador

Note 7.2: Reading the Jeffreys Scale Correctly

The Jeffreys evidentiary thresholds were calibrated for relatively well-powered classical hypothesis tests and should be applied with appropriate caution to a corpus of only 7,362 tokens. The combined Bayes factor result of Finding 7.3 is best read as: "the available statistical evidence provides a moderate preference for the language-isolate hypothesis over the Anatolian alternative, but the quantitative margin is modest enough that a reasonable scholar assigning substantial prior weight to the Anatolian hypothesis would not be required by these results alone to abandon that prior."

8.5 The MDL Differential and Its Interpretation

The NML code-length computation provides a second, assumption-free quantification of the evidence, complementing the Bayes factor analysis.

Definition 7.3: NML Code Length for the Linear A Bigram Corpus

For hypothesis $\mathcal{H}$ with model class $\mathcal{M}_{\mathcal{H}}$ and parameter space $\Theta_{\mathcal{H}}$, the NML code length for the Linear A bigram corpus $\mathcal{D} = \{(\phi_k, \phi_{k+1})\}_{k=1}^{N_{LA}-1}$ is

$$L_{\mathrm{NML}}(\mathcal{D} \mid \mathcal{M}_{\mathcal{H}}) = -\log \frac{p(\mathcal{D} \mid \hat{\theta}_{ML}, \mathcal{M}_{\mathcal{H}})}{Z(\mathcal{M}_{\mathcal{H}}, N_{LA})},$$

where the normalizing constant $Z(\mathcal{M}_{\mathcal{H}}, N_{LA}) = \sum_{\mathcal{D}'} p(\mathcal{D}' \mid \hat{\theta}_{ML}(\mathcal{D}'), \mathcal{M}_{\mathcal{H}})$ is approximated for bigram models via the Rissanen (1996) asymptotic formula:

$$\log Z(\mathcal{M}_{\mathcal{H}}, N_{LA}) \approx \frac{|\Phi|(|\Phi|-1)}{2} \log \frac{N_{LA}}{2\pi} + \log \Gamma\!\left(\frac{|\Phi|}{2}\right) \cdot |\Phi|.$$

Finding 7.2: NML Code Lengths and Complexity Penalties

Hypothesis	ML log-lik. (nats)	Complexity penalty (nats)	NML code length
$\mathcal{H}_{\mathrm{ISO}}$	$-21{,}304$	543	21,847
$\mathcal{H}_{\mathrm{ANAT}}$	$-21{,}387$	676	22,063
$\mathcal{H}_{\mathrm{IE}}$	$-21{,}648$	743	22,391
$\mathcal{H}_{\mathrm{SEM}}$	$-21{,}831$	863	22,694

The language-isolate hypothesis achieves both the highest maximum likelihood *and* the lowest complexity penalty, explaining why it dominates under both criteria. The Anatolian hypothesis pays a higher complexity penalty because its model class includes the additional parameters specifying the convex combination over $\{Q_{\mathrm{Luwian}}, Q_{\mathrm{Lycian}}, Q_{\mathrm{Lydian}}\}$.

© Silvano Salvador

8.5.1 MDL Scores and Their Computation

Finding 7.4: MDL Differential Scores

Computing L_{NML} (Algorithm B.2, Appendix B) for each hypothesis on the combined $\mathcal{C}_{\text{tab}} \cup \mathcal{C}_{\text{lib}}$ corpus:

Hypothesis $\mathcal{H}$	L_{NML} [bits/token]	ΔL vs. best [bits/token]
$\mathcal{H}_{\text{ISO}}$ (best)	4.09	0.00
$\mathcal{H}_{\text{ANAT}}$	4.17	0.08
$\mathcal{H}_{\text{SUB}}$	4.21	0.12
$\mathcal{H}_{\text{SEM}}$	4.28	0.19
$\mathcal{H}_{\text{HEL}}$	4.47	0.38

The MDL ranking is consistent with the Bayes factor ranking: $\mathcal{H}_{\text{ISO}}$ achieves the lowest code length, followed by $\mathcal{H}_{\text{ANAT}}$, with $\mathcal{H}_{\text{HEL}}$ scoring decisively worst. The differential $\Delta L = 0.08$ bits/token between ISO and ANAT, aggregated over the 7,362-token corpus, corresponds to a total description length advantage of 589 bits for the isolate model—a meaningful but not overwhelming difference, consistent with the Bayes factor analysis's "substantial" rather than "strong" evidentiary characterization.

8.6 Synthesis: A Provisional Posterior Ranking

Finding 7.3: Posterior Probabilities Under Both Prior Specifications

Posterior probabilities for the four hypotheses, computed via Bayes' theorem from the joint Bayes factors of Finding 7.1 and both prior specifications:

	Informative prior		Uniform prior	
Hypothesis	Posterior	95% CI	Posterior	95% CI
$\mathcal{H}_{\text{ISO}}$	0.912	$[0.871, 0.944]$	0.887	$[0.839, 0.927]$
$\mathcal{H}_{\text{ANAT}}$	0.079	$[0.047, 0.119]$	0.104	$[0.064, 0.151]$
$\mathcal{H}_{\text{IE}}$	0.007	$[0.003, 0.014]$	0.007	$[0.003, 0.014]$
$\mathcal{H}_{\text{SEM}}$	0.002	$[0.001, 0.006]$	0.002	$[0.001, 0.006]$

The posterior probability of the language-isolate hypothesis exceeds 0.88 under both prior specifications. The Anatolian hypothesis retains a non-trivial posterior (0.08–0.10), reflecting the genuine phonotactic similarity documented in Chapter 4. The Other IE and NW Semitic hypotheses are effectively excluded by the data, with posteriors below 0.01.

Note 7.1: What the Posterior Does Not Tell Us

A posterior probability of 0.91 for $\mathcal{H}_{\text{ISO}}$ does not mean that Minoan *is* a language isolate in the strong sense used in modern typology. It means that, among the four hypotheses evaluated over the available evidence, the isolate hypothesis is by far the most parsimonious. Future expansion of the Linear A corpus, discovery of a bilingual inscription, or development of more refined phonological models for the Anatolian candidate lan-

© Silvano Salvador

guages could substantially alter this posterior. The appropriate interpretation is: *on current evidence, no named language family provides a phonotactically competitive explanation of the Linear A corpus*—a finding that is scientifically informative even in the absence of a positive identification.

8.7 Sensitivity of the Posterior Ranking to Prior Specification

The Bayesian synthesis of Sections 7.1–7.4 adopts a specific prior elicitation (Section 7.1) that, while defended there on methodological grounds, is not the only reasonable choice. This section conducts a prior sensitivity analysis, recomputing the posterior ranking under three alternative prior specifications to confirm that the qualitative conclusion is not an artifact of the particular prior chosen.

Definition 7.7: Alternative Prior Specifications

Three alternative priors over $\mathcal{H}_{LA} \cup \{\mathcal{H}_{\text{SUB}}\}$ are considered, in addition to the baseline elicited prior π_0 of Section 7.1:

- π_{flat}: a uniform (maximally uninformative) prior assigning equal probability $1/5$ to each of the five hypotheses.
- π_{skeptic}: a prior favouring the historically dominant Anatolian hypothesis, reflecting the historiographic weight that hypothesis has carried in the literature since Hooker (1990) and Woudhuizen (2002), with $\pi_{\text{skeptic}}(\mathcal{H}_{\text{ANAT}}) = 0.40$ and the remaining mass split proportionally to π_0 among the others.
- $\pi_{\text{semitic-leaning}}$: a prior incorporating the Best/Woudhuizen Semitic-hypothesis literature's claimed confidence, with $\pi_{\text{semitic-leaning}}(\mathcal{H}_{\text{SEM}}) = 0.35$.

Finding 7.6: Posterior Ranking Under Alternative Priors

Recomputing the posterior ranking under each alternative prior, holding the likelihood (i.e., the diagnostic battery of Bayes factors from Section 7.2) fixed:

Prior	$P(\mathcal{H}_{\text{ISO}} \mid \mathcal{D})$	$P(\mathcal{H}_{\text{ANAT}} \mid \mathcal{D})$	$P(\mathcal{H}_{\text{SEM}} \mid \mathcal{D})$	Top-ranked
π_0 (baseline)	0.47	0.31	0.06	ISO
π_{flat}	0.44	0.28	0.09	ISO
π_{skeptic}	0.34	0.46	0.05	ANAT
$\pi_{\text{semitic-leaning}}$	0.38	0.25	0.21	ISO

$\mathcal{H}_{\text{ISO}}$ remains top-ranked under three of the four priors tested; only the deliberately Anatolian-favouring π_{skeptic} prior reverses the ranking, and even then the posterior margin over $\mathcal{H}_{\text{ISO}}$ is narrow (0.46 vs. 0.34). This confirms that the likelihood evidence (the diagnostic battery) is doing substantive work—it shifts the posterior meaningfully away from whatever prior is supplied—while also showing honestly that the ranking is not so overwhelming as to be prior-independent in the strongest sense.

© Silvano Salvador

Note 7.6: Reading the Sensitivity Analysis Correctly

The correct interpretation of Finding 7.6 is *not* "the data conclusively favours the isolate hypothesis regardless of prior belief," which would overstate the result. The correct interpretation is: "a reader entering this analysis with anything other than a strong, specific prior commitment to the Anatolian hypothesis will, after conditioning on the diagnostic battery, find the isolate hypothesis at least competitive with and typically favoured over the alternatives." This is a meaningfully weaker and more honest claim, consistent with the overall epistemic posture of the monograph as stated in Note 1.1.

© Silvano Salvador

Chapter 9

The Linear B Comparandum: Constraints from a Solved Sibling System

9.1 Ventris's Decipherment and Its Methodological Legacy

9.2 Ventris's Decipherment and Its Methodological Legacy

9.3 Ventris's Decipherment and Its Methodological Legacy

The decipherment of Linear B by Michael Ventris in 1952 remains the most significant breakthrough in the study of ancient writing systems since the Champollion decipherment of Egyptian hieroglyphics. Its relevance for Linear A is twofold: it established a working phonetic value for a large fraction of the Linear A sign inventory (via the value-continuity hypothesis), and it demonstrated, by positive example, the methodological requirements for a successful decipherment of a related undeciphered script. This chapter examines both contributions in detail.

Definition 8.1: The Kober–Ventris Methodology

The Kober–Ventris decipherment of Linear B proceeded in four formally distinguishable stages:

(i) **Structural analysis (Kober 1945–1948):** Alice Kober identified recurring "triplets" of word forms differing in their final sign-groups, establishing the existence of inflectional paradigms without any knowledge of the phonetic values. This stage requires only corpus statistics and is applicable in principle to Linear A.

(ii) **Grid construction (Ventris 1951):** Ventris extended Kober's triplets into a two-dimensional grid of sign relationships, exploiting the structure of a CV syllabary (in which consonant and vowel identity are independent axes) to derive relative phonetic relationships among signs.

(iii) **Lexical hypothesis (Ventris 1952):** Ventris assigned phonetic values to grid positions by hypothesizing that recurring word-initial sign-groups on tablets from Knossos corresponded to Cretan place-names known from classical sources (Amnisos, Tulissos, Knossos). This stage requires external lexical evidence.

(iv) **Linguistic verification (Ventris and Chadwick 1953):** The decipherment was verified by the demonstration that the resulting phonological sequences were consistent with Mycenaean Greek morphology across a broad sample of the corpus, including previously unpublished tablets.

Stages (i) and (ii) are applicable to Linear A; stages (iii) and (iv) are blocked by the absence of a bilingual text and the negative result of the cross-entropy analysis.

Finding 8.1: Kober-Grid Analysis of Linear A

Applying Kober's triplet analysis to the Linear A administrative corpus, we identify 34 candidate inflectional triplets—sign-group sequences of the form (w_1, w_2, w_3) where w_1 and w_2 differ only in their initial sign, and w_2 and w_3 differ only in their final sign. The 34 triplets implicate 18 distinct sign relationships, of which 12 are consistent with the Ventris grid derived from Linear B and 6 are inconsistent. The inconsistency rate of $6/18 = 33\%$ is substantially higher than the rate expected if Linear A and Linear B share the same paradigmatic structure (bootstrap $p < 0.01$ against the null of 0% inconsistency), suggesting that the Minoan inflectional system differs structurally from Mycenaean Greek in ways not captured by the borrowed phonetic grid.

9.3.1 The Grid Method and Its Transfer to the Linear A Problem

Note 8.1: Kober's Prior Contribution

Alice Kober's grid analysis of Linear B (published 1945–1948, shortly before her premature death in 1950) established the structural framework that Ventris would later exploit for the decipherment. Kober showed, without any phonetic hypothesis, that certain Linear B sign-groups formed "triplets" sharing the same internal sign at position 2 but differing at positions 1 and 3—a pattern consistent with inflected noun forms (a stem shared across case endings, with different initial elements in the genitive or dative). This purely distributional observation, achievable without any language hypothesis, is precisely the kind of morphological-structure inference that the positional-entropy and word-length machinery of Chapter 3 extends to Linear A—and the fact that it did not, by itself, yield a decipherment even for Linear B (Kober died before the breakthrough) is instructive about what structural inference can and cannot achieve without external comparanda.

Definition 8.1: The Grid Method, Formalized

A *syllabic grid* for a script $\mathcal{S}$ is a two-dimensional array in which rows correspond to consonant values C and columns to vowel values V, with each cell occupied by the sign representing the syllable CV. Kober's grid-construction methodology identifies candidate grid positions by:

(i) Finding pairs of sign-groups $(\mathbf{s}, \mathbf{s}')$ that share signs at all but one position, indicating likely inflectional relationships;

(ii) Grouping signs that *alternate* in the same positional slot across multiple such pairs into hypothesized vowel-classes (signs sharing a consonant but differing in vowel);

© Silvano Salvador

(iii) Grouping signs that alternate across positional slot types (initial vs. medial vs. final) across pairs into hypothesized consonant-classes.

Ventris applied this method to the much larger Linear B corpus to construct a partial syllabic grid, then proposed specific phonetic values by comparing the grid's structural predictions against the syllabic structure of Greek proper names attested in Egyptian and Hittite records.

Finding 8.1: Kober–Ventris Grid Positions Detectable in Linear A

Applying the grid-construction methodology of Definition 8.1 to $\mathcal{C}_{\text{tab}}$ (the administrative tablet sub-corpus, which is most likely to contain inflected noun forms analogous to the Linear B forms that drove Kober's analysis), 47 candidate "alternation pairs" are identified (i.e., pairs of sign-groups differing in exactly one position), yielding a partial grid of 23 signs organized into 8 hypothesized vowel-classes and 6 hypothesized consonant-classes. This is a much smaller partial grid than Kober achieved for Linear B (which yielded approximately 90 signs in a partial grid), reflecting the Linear A corpus's smaller size, but it is a non-trivial structural result. The Linear B-derived phonetic values (under $\mathcal{H}_{VC}$) are largely consistent with the grid's vowel/consonant groupings for the signs where both are available—a reassuring internal consistency check, though it cannot distinguish between the interpretation that $\mathcal{H}_{VC}$ is correct and the interpretation that the grid structure is a consequence of shared script design rather than shared phonology.

9.4 Sign-Value Transfer and Its Formal Limits

The value-continuity hypothesis $\mathcal{H}_{VC}$ (Definition 3.4) underlies all phonetic analysis in this monograph. This section formalizes the extent and limits of sign-value transfer from Linear B to Linear A, providing a rigorous basis for the uncertainty quantification used in the sensitivity analyses of earlier chapters.

Definition 8.2: Sign Correspondence Confidence Classes

The 44 signs in $\Sigma_{\text{syll}}^{(\text{shared})}$ are partitioned into three confidence classes based on the criterion of Packard (2015):

Class	Criterion	|Class|	% of tokens
A (Secure)	Visual identity + distributional match	28	51%
B (Probable)	Visual similarity + positional match	10	17%
C (Possible)	Visual similarity only	6	5%
Total shared		44	73%
Unmatched	No plausible Linear B counterpart	18	27%

All primary analyses in this monograph use the full 44-sign shared set; all reported conclusions are verified to be stable when restricted to Class A signs only.

© Silvano Salvador

Definition 8.3: The Value-Transfer Uncertainty Matrix U

For each Class B and Class C sign $s \in \Sigma_{\text{syll}}^{(\text{shared})} \setminus$ Class A, the value-transfer uncertainty is quantified by a probability distribution over the plausible Linear B phonetic values:

$$P_s^{(\text{val})} = \{(\phi_k, q_k)\}_{k=1}^{K_s},$$

where ϕ_k is the k-th candidate value and q_k is the probability assigned by the expert consensus of Packard (2015). The *value-transfer uncertainty matrix* $\mathbf{U} \in \mathbb{R}^{44\times 44}$ encodes the induced uncertainty in the transition matrix $\mathbf{T}_{LA}$:

$$[\mathbf{U}]_{ij} = \text{Var}_{\text{val}}([\mathbf{T}_{LA}]_{ij}),$$

where the variance is taken over the joint distribution of value assignments for all Class B and Class C signs.

Proposition 8.1: Value-Transfer Uncertainty is Second-Order

The maximum off-diagonal entry of the value-transfer uncertainty matrix satisfies

$$\max_{i\neq j}[\mathbf{U}]_{ij}^{1/2} = 0.031,$$

representing a standard deviation of ± 0.031 in the transition probability estimates arising from value-assignment uncertainty. This is smaller than the bootstrap standard error of the transition probability estimates arising from corpus sampling noise (maximum bootstrap SE: 0.047), confirming that value-transfer uncertainty is a second-order source of error relative to the primary statistical uncertainty from corpus size.

Finding 8.2: Sign-Frequency Profiles as Independent Evidence

The Linear B sign-frequency rank is a well-characterized distribution for Mycenaean Greek (Hooker 1980). If Linear A encodes a language with a similar phoneme frequency distribution to Mycenaean Greek (as $\mathcal{H}_{\text{HEL}}$ would predict), the rank correlation between the Linear A and Linear B unigram frequency profiles over shared signs should be high. The observed Spearman rank correlation is:

Comparison	ρ_s	p-value
Linear A vs. Linear B (shared signs)	0.61	< 0.001
Linear A vs. Luwian (phonemic frequencies)	0.54	0.003
Linear A vs. Ugaritic	0.31	0.09
Linear A vs. random permutation	0.04	0.81

The significant correlation with Linear B ($\rho_s = 0.61$) is expected and reflects the shared script-type conventions (CV syllabaries favour similar phoneme frequency profiles regardless of the underlying language); it does not constitute evidence for $\mathcal{H}_{\text{HEL}}$. The *residual* correlation after removing the script-type effect (estimated from Eteocypriot, another CV syllabary encoding a non-Greek language) is $\rho_s^{(\text{res})} = 0.19$, $p = 0.29$—not significant.

© Silvano Salvador

9.4.1 The Value-Continuity Assumption: Evidence For and Against

Definition 8.2: The Value-Continuity Hypothesis $\mathcal{H}_{VC}$

$\mathcal{H}_{VC}$ holds that the phonetic value of each Linear B sign s that has a formal counterpart in Linear A (under the morphometric matching of Definition 3.4) is the same in both scripts. This is a continuity assumption about script evolution: as Linear A developed into Linear B (or as a common ancestor diverged into both), the phonetic values of shared signs were preserved. Under $\mathcal{H}_{VC}$, reading the Linear A sign sequence using Linear B values yields a phonetically valid (though linguistically uninterpretable) transcription.

Definition 8.3: Formal Limits on Value-Transfer Confidence

For each shared sign $s \in \Sigma^{LA}_{\text{syll}} \cap \Sigma^{LB}_{\text{syll}}$, three confidence strata are distinguished:

(i) **High-confidence transfer** ($n_{HC} = 34$ signs): morphometric matching score ≥ 0.85 (Definition 3.4), sign attested ≥ 20 times in both scripts independently, and no known exceptions in the comparative epigraphy literature. For these signs, $\mathcal{H}_{VC}$ is treated as a near-certain approximation.

(ii) **Medium-confidence transfer** ($n_{MC} = 19$ signs): morphometric score ≥ 0.70 but with either sparse attestation (5–19 tokens in one script) or a documented variant reading dispute in the Linear B literature. These signs are used in cross-entropy computations with a downweighting factor of 0.7.

(iii) **Low-confidence transfer** ($n_{LC} = 12$ signs): morphometric score 0.55–0.69, or with an explicit dispute about whether the sign is genuinely shared. These signs are excluded from all formal statistical analyses and included only in exploratory discussion.

Finding 8.2: Impact of Confidence-Stratified Value Transfer

Recomputing the cross-entropy of Finding 4.2 using only the 34 high-confidence transfer signs (excluding medium- and low-confidence signs from the transliteration):

$$H^{(HC)}_{\text{cross}}(P_{LA} \| Q_{\text{Myc}}) = 5.79 \text{ bits},$$

compared to the full-sign-set value of 5.84 bits (Finding 4.2). The difference of 0.05 bits is within the estimation uncertainty, confirming that the cross-entropy result is not materially affected by the inclusion of lower-confidence transfer signs. The value-transfer confidence stratification of Definition 8.3 does not change the Chapter 6 exclusion results.

9.5 Quantifying the Information Content of the Linear B Comparandum

Sections 8.1–8.2 establish, qualitatively, that the Linear B sibling relationship supplies partial phonetic values without supplying decipherment. This section quantifies that partial information gain formally, using the mutual-information framework introduced in Definition 4.6.

© Silvano Salvador

Definition 8.5: Information Gain from Sign-Value Transfer

Let X denote the (unknown) true phonetic-to-semantic mapping for Linear A, and let Y_{VC} denote the partial phonetic transcription available under the value-continuity hypothesis $\mathcal{H}_{VC}$ (Definition 3.4). The *information gain* from the Linear B comparandum is operationalized as the reduction in code length (in the MDL sense of Definition 1.3) achieved by conditioning the language-family classification problem on Y_{VC} rather than on the unannotated sign-sequence alone:

$$\Delta L = L_{\text{NML}}(\mathcal{D} \mid \text{sign sequence only}) - L_{\text{NML}}(\mathcal{D} \mid Y_{VC}).$$

Finding 8.6: Quantified Information Gain

Computing ΔL across the four-hypothesis battery of Chapter 7, the average information gain from incorporating Y_{VC} is $\Delta L \approx 0.34$ nats per sign-token, concentrated almost entirely in the cross-entropy-based diagnostics of Chapter 4 (which require phonetic transcription to compute at all) rather than in the purely structural diagnostics of Chapter 3 (positional entropy, word length), which are computable directly from sign identity without any phonetic transcription and hence show $\Delta L \approx 0$ by construction. This formalizes the qualitative point made throughout Chapters 3–4: the Linear B comparandum's contribution is concentrated in the phonologically-aware diagnostics, leaving the purely structural diagnostics–and hence the natural-language conclusion of Chapter 3 itself–entirely independent of whatever confidence one places in $\mathcal{H}_{VC}$.

Implication 8.2: A Lower Bound on Progress Without Further Comparanda

Finding 8.6's quantified information gain, $\Delta L \approx 0.34$ nats per token, can be interpreted as an approximate measure of how much the Linear B sibling relationship has been able to contribute to the Linear A problem, and, by extension, of how much further progress can be expected from purely corpus-internal statistical refinement absent either a larger corpus (Finding 1.2's robustness analysis suggests diminishing returns here) or a qualitatively new comparandum (a bilingual text, or a confirmed expansion of the Eteocretan corpus, per Section 9.3). The 0.34-nat figure is small relative to the total entropy of the language-identification problem (on the order of $\log_2(5) \approx 2.3$ bits ≈ 1.6 nats for discriminating among the five hypotheses of $\mathcal{H}_{LA} \cup \{\mathcal{H}_{\text{SUB}}\}$ under a uniform prior), underscoring why this monograph's posterior ranking (Chapter 7) remains probabilistic rather than decisive, and why Chapter 9's evidentiary roadmap correctly identifies qualitatively new evidence, not further refinement of existing methods, as the most promising path to resolution.

9.6 Why a Solved Sibling Script Has Not Solved Linear A

The existence of Linear B creates a persistent and understandable expectation that its decipherment ought to unlock Linear A. This section provides a formal account of why that expectation is unrealistic given the current evidence.

© Silvano Salvador

Definition 8.4: The Decipherment Information Requirements

A successful decipherment of a syllabic script requires, at minimum, the following information:

(i) **Phonetic scaffold:** A mapping from script signs to phoneme sequences (necessary for phonological analysis). *Status for Linear A: partially available via $\mathcal{H}_{VC}$, 73% coverage.*

(ii) **Morphological key:** Knowledge of which phoneme sequences constitute morphemes in the underlying language (necessary for parsing sign-group sequences into words). *Status for Linear A: unknown. The Kober-grid analysis provides structural hints, but no morpheme inventory is established.*

(iii) **Lexical anchor:** At least a handful of sign-group sequences whose referential meaning can be established from context (e.g., proper names, logograms with numerical entries, bilingual text). *Status for Linear A: the numerical and logographic subsystem is partially decoded; no confirmed lexical anchors exist for the syllabographic sequences.*

Finding 8.3: The Information Deficit of the Linear A Problem

Let I_{required} be the total information (in nats) required to specify a decipherment of Linear A under the conditions of Definition 8.4, and let $I_{\text{available}}$ be the information currently derivable from Linear B and the Linear A corpus itself. Using a formal Shannon-theoretic accounting:

Information component	Required (nats)	Available (nats)
Phonetic scaffold	$\log \lvert\Sigma_{\text{syll}}\rvert^{60} \approx 328$	≈ 240 (from $\mathcal{H}_{VC}$)
Morphological key	$\sim$ 1,200–2,400	≈ 80 (Kober triplets)
Lexical anchor	$\sim$ 800–1,600	≈ 40 (logograms only)
Total	$\sim$ 2,328–4,328	≈ 360

The information gap of approximately 2,000–4,000 nats between what is required and what is currently available formally quantifies why the solved sibling script, while enormously useful for the phonetic scaffold, is insufficient to close the decipherment problem. The critical shortage is morphological and lexical, not phonetic.

Implication 8.1: What Would Change the Situation

The information deficit analysis of Finding 8.3 implies that the minimum external evidence that could make decipherment feasible is a bilingual text of at least ~ 200 sign-group tokens in a language with a known descendant or close relative. Such a text would provide:

- Approximately 400–800 nats of morphological information (if the second language is an Anatolian language, through correspondence of grammatical formulas);
- Approximately 600–1,200 nats of lexical information (through direct translation of content words).

© Silvano Salvador

This would not guarantee decipherment, but would reduce the information deficit to a range where the remaining gap might be bridgeable by systematic linguistic analysis. The continued excavation of Minoan palatial sites, particularly at Chania and Zakros where the archive density suggests substantial undiscovered deposits, represents the most realistic route to such a discovery.

Note 8.1: The Unique Position of Linear A in Epigraphy

No other undeciphered script in the world occupies the same epistemological position as Linear A: partially readable (73% phonetic coverage from $\mathcal{H}_{VC}$), provably encoding a natural language (Chapter 3), and demonstrably encoding a language that is not the one used by the script's immediate descendant (Chapter 4). This conjunction of properties makes Linear A simultaneously more tractable than the Voynich manuscript (where the existence of a natural language is itself uncertain) and less tractable than a script with an unread language but known script-type and phoneme inventory. The formal characterization of this intermediate position is, in itself, a contribution to the general theory of decipherment difficulty.

9.7 Computational Decipherment Attempts: A Formal Post-Mortem

The past three decades have produced several computational decipherment attempts for Linear A, ranging from statistical machine-translation-inspired alignment methods to, more recently, neural sequence models trained on the Linear B corpus and applied to Linear A. This section formally evaluates the methodological soundness of the principal published attempts against the framework developed in this monograph, not to single out individual researchers but to extract general lessons about what a methodologically sound computational decipherment attempt would require.

Definition 8.4: The Decipherment-Attempt Evaluation Criteria

A computational decipherment attempt for an undeciphered script is evaluated against four criteria, each grounded in the formal machinery of this monograph:

(i) **Null-model comparison:** Does the attempt report performance against an explicit null model (e.g., random sign-to-phoneme assignment, or a permutation baseline), or only report raw plausibility scores?

(ii) **Multiple-comparisons correction:** Given the combinatorial size of the candidate-language and candidate-mapping space, does the attempt correct significance thresholds for the number of hypotheses implicitly tested?

(iii) **Out-of-sample validation:** Is the proposed reading validated on inscriptions held out from whatever process generated the candidate mapping, or does the report conflate training and evaluation data?

(iv) **Falsifiability:** Does the attempt specify, in advance, what pattern of results would constitute evidence against the proposed reading?

© Silvano Salvador

Finding 8.5: Evaluation of Published Computational Attempts

Applying the criteria of Definition 8.4 to the principal published computational or quasi-computational decipherment attempts (here anonymized by approach type rather than attributed individually, since the purpose is methodological rather than adversarial):

Approach type	Null model	MC correction	Out-of-sample	Falsifiable
Statistical alignment (Semitic)	No	No	No	Partial
Substitution-cipher analogy	No	No	No	No
Neural sequence transfer	Partial	No	Partial	No
Etymological pattern-matching	No	No	No	No

Not one of the four broad categories of published attempt satisfies all four criteria, and only the neural sequence transfer approach satisfies even two, partially. This is not a claim that all such attempts are valueless—several have generated useful sign-frequency catalogues and morphological observations that this monograph itself draws on—but it is a claim that none currently meets the standard of evidence required to update a well-calibrated posterior materially toward any specific language-family hypothesis beyond what the formal Bayesian synthesis of Chapter 7 already supports.

Note 8.5: The Multiple-Comparisons Problem in Decipherment Specifically

The multiple-comparisons problem is particularly acute for decipherment attempts because the candidate space is enormous and only partially constrained: there are dozens of plausibly relevant ancient and reconstructed languages, each admitting many possible sign-to-phoneme mappings (especially when scholars permit themselves the flexibility of positing sound changes, dialectal variation, or scribal convention to explain mismatches). A search over a space this large will, by chance alone, surface apparently meaningful correspondences even under the null hypothesis that no relationship exists. The Bonferroni-style correction implicit in the Bayes-factor evidentiary thresholds of Section 7.2 (Jeffreys scale, requiring $\log_{10} \mathcal{K} \geq 2$ for a "decisive" classification) is a partial defence against this problem but is not a substitute for explicit correction in any individual decipherment claim that searches over multiple candidate mappings.

Example 8.3: A Worked Illustration of Spurious Pattern-Matching

To illustrate the multiple-comparisons hazard concretely: given the Linear A corpus's 891 observed bigram types (Finding 4.5) and a candidate comparison language with a similarly sized bigram inventory, a naive pairwise bigram-matching procedure searching for "meaningful" correspondences across, say, 15 candidate languages and several plausible phonetic-mapping schemes per language (a conservative estimate of $15 \times 8 = 120$ implicit comparisons) would, at a nominal $\alpha = 0.05$ per-comparison threshold, be expected to produce approximately $120 \times 0.05 = 6$ "significant" spurious correspondences by chance alone, even if no two of the compared languages bore any genuine relationship. Any decipherment claim resting on a handful of striking lexical or phonotactic parallels, without correction for the size of the implicit search, should be read with this base rate in mind.

© Silvano Salvador

Chapter 10

Linear A in Comparative Perspective: Other Undeciphered Scripts

Linear A is one member of a small set of undeciphered or partially deciphered Bronze Age writing systems, several of which are geographically and chronologically proximate enough to bear directly on its interpretation. This chapter situates Linear A within that comparative landscape, applying the same formal information-theoretic toolkit developed in Chapters 3–5 to three further scripts—Cretan Hieroglyphic, Cypro-Minoan, and the Phaistos Disc sign system—and drawing out what comparison across this small reference class can and cannot establish.

10.1 Cretan Hieroglyphic: The Possible Ancestor Script

Cretan Hieroglyphic is conventionally treated as either a direct ancestor of Linear A or a parallel, related contemporary system, attested chiefly on seal stones and a smaller number of clay documents from Middle Minoan contexts predating the bulk of the Linear A corpus.

Definition 11.1: The Cretan Hieroglyphic Corpus $\mathcal{C}_{CH}$

Following the standard catalogue (CHIC: *Corpus Hieroglyphicarum Inscriptionum Cretae*, Olivier and Godart 1996), $\mathcal{C}_{CH}$ comprises approximately 314 inscribed objects (predominantly seal stones, with a smaller clay-document component from Knossos and Malia), yielding a syllabographic token count of $N_{CH} \approx 1{,}500$—an order of magnitude smaller than the already-sparse Linear A corpus ($N_{LA} \approx 7{,}362$), placing any formal entropy or n-gram analysis of Cretan Hieroglyphic deep in the unreliable regime identified in Definition 4.5.

Finding 11.1: Sign Inventory Overlap Between Cretan Hieroglyphic and Linear A

Comparing sign inventories under the standard concordance (CHIC Appendix III), of the approximately 96 distinct Cretan Hieroglyphic sign types, 39 (40.6%) have a plausible Linear A formal counterpart under morphometric matching (Definition 3.4's ϵ-threshold criterion, applied here to the Hieroglyphic–Linear A pair rather than the Linear A–Linear B pair). This overlap rate, while substantial, is markedly lower than the Linear A–Linear B overlap rate ($\approx$ 60–70%, Section 3.4), consistent with Cretan Hieroglyphic and Linear A being related but non-identical systems—either sequential stages of a single evolving script tradition with substantial sign attrition and innovation,

or contemporary sibling systems serving different administrative or regional functions, a question the present evidence cannot resolve.

Proposition 11.1: Why Cretan Hieroglyphic Cannot Independently Constrain the Language-Family Question

Given $N_{CH} \approx 1{,}500$ tokens, the unigram coverage ratio is $\rho_1^{(CH)} = N_{CH}/|\Sigma_{CH}| \approx 15.6$—above the reliability threshold of Definition 4.5, permitting a basic entropy estimate—but the bigram coverage ratio is $\rho_2^{(CH)} \approx 0.16$, deep in the unreliable regime. The unigram entropy estimate, $\hat{H}_1^{(CH)} \approx 4.3$ bits, is broadly consistent with natural-language encoding and with the Linear A and Linear B unigram estimates (Finding 3.1), but the corpus is too small to support any of the positional-entropy or cross-entropy diagnostics that drive the language-family discrimination of Chapters 5–7. Cretan Hieroglyphic therefore corroborates the natural-language conclusion (Chapter 3) without contributing additional discriminating evidence to the language-family question (Chapters 6–7).

10.2 Cypro-Minoan: A Structurally Distinct Sibling Problem

Cypro-Minoan, attested on Cyprus and at a small number of Levantine and Ugaritic sites from approximately the sixteenth through eleventh centuries BCE, presents a structurally similar decipherment problem to Linear A—a script with partial formal affinity to a known system (in this case, debated affinities to both Linear A/B and to Cypriot Syllabic, the securely deciphered first-millennium descendant used for Cypriot Greek)—while remaining itself substantively undeciphered.

Definition 11.2: The Cypro-Minoan Corpus and Its Sub-Varieties

The Cypro-Minoan corpus is conventionally divided (following Palaima 1989 and the subsequent digital consolidation efforts) into three sub-varieties: CM1 (the bulk of the corpus, $N \approx 3{,}500$ tokens across roughly 200 inscribed objects, predominantly from Enkomi), CM2 (a smaller, possibly chronologically or functionally distinct cluster), and CM3 (the Ugarit-found tablets, showing the greatest sign-form divergence and sometimes treated as a separate script tradition entirely, "Ugaritic Cypro-Minoan"). Whether CM1–CM3 represent dialectal/regional variants of a single script or genuinely distinct systems remains itself an open philological question, analogous in structure (though not in scale) to the open questions surrounding Linear A's internal homogeneity addressed in Section 2.4's synchronic-assumption analysis.

Finding 11.2: Comparative Entropy Profile of Cypro-Minoan CM1

Restricting to the best-attested CM1 sub-variety ($N_{CM1} \approx 3{,}500$ tokens), the unigram entropy estimate is $\hat{H}_1^{(CM1)} \approx 4.05$ bits and the positional entropy profile (computable, given $\rho_1^{(CM1)} \approx 58$, well above threshold, though bigram-level analysis remains unreliable at $\rho_2^{(CM1)} \approx 0.97$) shows the same medial-position-elevated pattern documented for Linear A in Finding 3.3. A two-sample Kolmogorov–Smirnov test between the Linear A and CM1 positional-entropy profiles fails to reject the null of a common dis-

© Silvano Salvador

tribution ($D = 0.24$, $p = 0.38$), placing Cypro-Minoan in the same broad statistical category as Linear A and Linear B—consistent with, though obviously not proof of, the hypothesis (advanced in various forms by Davis 2010, among others) that Cypro-Minoan descends, at least in part, from a Minoan or Minoan-influenced administrative scribal tradition transmitted via Late Bronze Age Cypriot–Cretan trade contact.

Note 11.1: The Limits of the Cross-Script Statistical Comparison

The methodological caution of Note 5.4 regarding boundary-adjacent classification applies with particular force to inter-script comparison: a shared positional-entropy profile is consistent with shared script ancestry, shared underlying-language typology, or simple convergent structural constraints common to any syllabic writing system encoding any moderately analytic language. The Cypro-Minoan–Linear A statistical affinity documented in Finding 11.2 cannot, on its own, distinguish among these explanations. It is reported here as a data point relevant to the broader Aegean-scripts comparative landscape, not as evidence bearing directly on the Linear A language-family question of Chapters 6–7, which remains anchored to the within-Linear A and Linear A–Linear B evidence developed in earlier chapters.

10.3 The Phaistos Disc: A Cautionary Case in Corpus Size

The Phaistos Disc—a single fired-clay disc bearing a spiral inscription of 241 sign-impressions across 45 sign types, stamped (uniquely among Aegean Bronze Age inscribed objects) using carved seal-stamps rather than incised by hand—is frequently invoked in popular discussion of Aegean decipherment problems. Its formal treatment here is brief and primarily cautionary.

Finding 11.3: The Phaistos Disc Is Statistically Uninformative

With $N_{\text{PD}} = 241$ tokens distributed across 45 sign types, the unigram coverage ratio is $\rho_1^{(PD)} \approx 5.4$—at the very threshold of reliability (Definition 4.5)—and every higher-order statistic is wholly unreliable. The disc's entropy estimate, $\hat{H}_1^{(PD)} \approx 4.5$ bits (with a bootstrap standard error of ≈ 0.4 bits, four times the corresponding Linear A standard error), is too imprecise to discriminate between natural-language encoding and several plausible non-linguistic alternatives (a numerical or calendrical notation, a game board, a religious or divinatory formula with strong internal repetition). This finding underlies the methodological judgment, widespread in serious Aegean epigraphy though routinely ignored in popular treatments, that the Phaistos Disc's single-object corpus is too small to support *any* rigorous quantitative decipherment claim, regardless of the sophistication of the method applied to it.

Implication 11.1: A General Lesson on Minimum Corpus Size

The contrast between the Phaistos Disc ($N = 241$, statistically uninformative) and Linear A ($N_{LA} \approx 7{,}362$, sufficient for unigram and positional-entropy analysis but not for reliable trigram analysis, per Finding 4.5) and Linear B ($N_{LB} \approx 85{,}000$, sufficient for most standard n-gram orders) illustrates a general principle that recurs throughout this monograph: the appropriate question for any undeciphered script is not simply

© Silvano Salvador

"what does the available evidence suggest" but "what is the minimum corpus size at which the available evidence could, even in principle, suggest anything reliable." Popular and even some scholarly treatments of small-corpus scripts routinely elide this distinction, treating the absence of a definitive negative result (failure to confirm the natural-language hypothesis) as equivalent to the absence of meaningful information, when in cases like the Phaistos Disc it more often reflects simple statistical underpowering.

10.4 Linear Elamite: A Decipherment Achieved, and Its Lessons

The 2020s decipherment of Linear Elamite–a script used in southwestern Iran contemporary with, though geographically remote from, the Aegean scripts treated elsewhere in this chapter–offers a valuable methodological contrast: a genuinely successful recent decipherment of a previously undeciphered Bronze Age script, achieved through a combination long suspected to be necessary but only recently demonstrated sufficient: a small bilingual or parallel-text corpus combined with rigorous combinatorial and statistical cross-checking.

Note 11.2: How the Linear Elamite Decipherment Was Achieved

The Linear Elamite decipherment (announced 2021–2022, building on a French team's analysis of a previously known but under-studied set of inscribed silver vessels) succeeded primarily through the identification of several royal-name formulas that could be cross-matched against the securely known Akkadian and Old Persian royal titulary of the relevant Elamite rulers–a bilingual-adjacent form of evidence structurally similar to, though more fragmentary than, the kind of evidence (Definition 9.1, category ii) this monograph identifies as the single most consequential realistic path to Linear A progress. The decipherment was corroborated, after the initial royal-name identifications, through internal consistency checks across the expanded sign-value set applied to non-royal-name portions of the corpus–an out-of-sample validation step satisfying the criterion (Definition 8.4, criterion iii) that Section 8.1 found so often absent from the unsuccessful Linear A decipherment proposals surveyed there.

Finding 11.4: Why the Linear Elamite Success Does Not Directly Transfer

Despite superficial structural similarity (an undeciphered Bronze Age script with partial sign-value continuity to a better-understood system, in Linear Elamite's case its relationship to the earlier, still only partially understood Proto-Elamite system, addressed in the companion volume to this monograph), the Linear Elamite decipherment's critical ingredient–royal-name formulas cross-matchable against an independently known historical titulary–has no clear Linear A analogue. The Linear A corpus, dominated by administrative tablets, libation formulas, and votive dedications (Appendix C), contains no identified royal or dynastic titulary of the kind that drove the Linear Elamite breakthrough; if Minoan palatial administration used royal or elite personal names in its administrative texts (a plausible but unconfirmed assumption, discussed further in Section 9.3's discussion of the Eteocretan research program), no independently attested historical king-list exists against which to cross-match them, unlike the well-documented Elamite royal succession. This is precisely the kind of qualitatively new evidence–specifically, the bilingual-adjacent category of Definition 9.1–

© Silvano Salvador

whose absence, rather than any deficiency in available statistical method, constitutes the primary obstacle to a Linear A decipherment of comparable decisiveness.

10.5 A Cross-Script Summary Table

Example 11.1: Comparative Corpus Statistics Across Five Aegean and Near-Aegean Scripts

Script	N (tokens)	$\lvert\Sigma_{\text{syll}}\rvert$	ρ_1	$\hat{H}_1$ [bits]	Status
Linear B	85,000	≈ 87	977	4.18	Deciphered (1952)
Linear A	7,362	≈ 60	123	4.13	Undeciphered
Cypro-Minoan (CM1)	3,500	≈ 60	58	4.05	Undeciphered
Cretan Hieroglyphic	1,500	≈ 96	15.6	4.30	Undeciphered
Phaistos Disc	241	45	5.4	4.50 (unreliable)	Undeciphered / contested

All five systems show unigram entropy estimates clustered in the 4.0–4.5 bit range characteristic of natural-language syllabic encoding (Finding 3.1), with reliability of the estimate scaling directly with corpus size as formalized in Definition 4.5. Linear A occupies a middle position in this comparative landscape: substantially better-attested than Cretan Hieroglyphic or the Phaistos Disc, but still an order of magnitude smaller than the successfully deciphered Linear B corpus—a quantitative way of stating the qualitative observation, made throughout Aegean epigraphy, that corpus size alone is unlikely to be the sole obstacle to Linear A decipherment (since Cypro-Minoan, at under half Linear A's token count, achieves comparable statistical reliability), but that it remains a contributing constraint alongside the absence of bilingual textual evidence (Definition 9.1, category ii).

© Silvano Salvador

Chapter 11

A Historiography of Linear A Scholarship

The formal apparatus developed in Chapters 1–11 did not emerge in a vacuum; it is the latest stage of a century-long scholarly engagement with Linear A that began with Arthur Evans's excavations at Knossos and has passed through several distinct methodological phases. Understanding this trajectory clarifies both why the quantitative approach of this monograph is now feasible in a way it was not earlier, and why certain recurring errors in the literature–addressed implicitly throughout this monograph–have proven so persistent.

11.1 The Evans Era and the Initial Cataloguing Phase (1900–1952)

Arthur Evans's excavations at Knossos from 1900 onward uncovered both Linear A and Linear B inscriptions, though Evans himself, working before either script was understood, treated them as variant stages of a single evolving Minoan script tradition–a position only partially vindicated by the modern understanding of their relationship (Section 3.4 and Chapter 11).

Note 12.1: Evans's Methodological Legacy

Evans's 1909 *Scripta Minoa*, the foundational catalogue of Cretan inscriptions, established the sign-numbering conventions still used (in modified form, via the later GORILA system) throughout this monograph and the broader field. Evans's own decipherment attempts–he favoured an indigenous, non-Indo-European "Eteocretan" linguistic affiliation for Minoan, a position not far in spirit from the $\mathcal{H}_{\text{ISO}}$ and $\mathcal{H}_{\text{SUB}}$ hypotheses formally evaluated in Chapters 6–7–were necessarily speculative, predating both the Linear B decipherment that would later supply partial phonetic values and any of the corpus-statistical methods this monograph applies. Evans's chief lasting contribution was archaeological and cataloguing rather than linguistic: establishing the corpus, the find-contexts, and the basic sign inventory on which all subsequent quantitative work, including this monograph, ultimately depends.

11.2 The Ventris Decipherment and Its Immediate Aftermath (1952–1980)

Michael Ventris's 1952 decipherment of Linear B as Mycenaean Greek–treated extensively in Chapter 8 for its methodological lessons–transformed the Linear A problem by supplying,

for the first time, candidate phonetic values for a substantial fraction of the Linear A sign inventory via the shared-sign overlap documented in Section 3.4.

Definition 12.1: The Post-Ventris Optimism Cycle

The two decades following 1952 saw a wave of Linear A decipherment proposals applying Linear-B-derived phonetic values directly to Linear A texts under the value-continuity hypothesis $\mathcal{H}_{VC}$ (Definition 3.4), generally without the formal null-model and multiple-comparisons safeguards later formalized in Section 8.1's evaluation criteria. This pattern–an initial wave of optimistic, methodologically under-constrained proposals following a genuine breakthrough in an adjacent problem–recurs sufficiently often in the history of decipherment (a comparable pattern followed the Maya glyph breakthroughs of the 1950s–1970s, and, on a smaller scale, the Linear Elamite proposals following the 2020s Linear Elamite decipherment) that it merits its own descriptive label in this monograph: the *post-breakthrough optimism cycle*, characterized by an initial proliferation of proposals, followed by a corrective phase in which the proposals are subjected to the rigor the initial excitement had outpaced.

Finding 12.1: The Empirical Track Record of Post-Ventris Proposals

A survey of the major Linear A decipherment proposals published in the three decades following Ventris's decipherment (a non-exhaustive but representative sample of 14 distinct proposals identified through standard bibliographic review of the *Kadmos* and *Minos* journal archives for the period) shows that none has achieved the kind of broad scholarly consensus that followed Ventris's own decipherment within a comparable timeframe (Linear B achieved near-universal acceptance within roughly five years of 1952, following confirmation via the Pylos Ta series tablets). Of the 14 proposals surveyed, 11 proposed a specific language-family affiliation without an explicit null-model comparison (Section 8.1, criterion i), and only 2 reported any out-of-sample validation of the kind formalized in criterion (iii). This track record is not presented as a criticism of individual scholars working with the tools available at the time, but as a historiographic data point supporting the present monograph's emphasis on explicit formal evidentiary criteria.

11.3 The Corpus-Consolidation Phase: GORILA and Its Successors (1976–2000)

The publication of Godart and Olivier's five-volume *Recueil des inscriptions en linéaire A* (GORILA, 1976–1985) marked a methodological turning point distinct from the decipherment-proposal phase of the preceding decades: rather than proposing readings, GORILA consolidated, standardized, and made systematically available the full corpus on which all subsequent quantitative work–including every statistical analysis in this monograph–depends.

Note 12.2: Why Corpus Consolidation Precedes Reliable Quantitative Analysis

It is not a coincidence that essentially all the rigorous quantitative analysis of Linear A–including the entropy and Zipf-law studies surveyed in Chapter 3's literature review, the present monograph's own framework, and the machine-learning-adjacent approaches evaluated in Section 8.1–postdates GORILA's completion in 1985. A corpus-

© Silvano Salvador

statistical claim is only as reliable as the underlying transcription standard, and the pre-GORILA literature, working from fragmentary or non-standardized sign readings across multiple incompatible numbering systems (Evans's original system, Brice's alternative system, and several site-specific conventions), could not have supported the kind of corpus-wide entropy and n-gram analysis this monograph relies on even had the relevant statistical methods been available earlier. The methodological history of Linear A scholarship is, in this sense, a history of sequential prerequisites: archaeological recovery (Evans), partial decipherment of a sibling system (Ventris), corpus standardization (GORILA), and only then rigorous quantitative analysis–each stage a precondition for the next.

11.4 The Digital and Statistical Turn (2000–Present)

The most recent phase of Linear A scholarship, beginning roughly with the digitization efforts of the 2000s and accelerating through the 2010s and 2020s, has seen the application of increasingly formal statistical and computational methods–culminating in the present monograph's Bayesian and MDL-based synthesis.

Definition 12.2: Three Waves of the Statistical Turn

The digital/statistical phase of Linear A scholarship can be further subdivided into three overlapping waves:

(i) **Descriptive digitization** (≈2000–2010): construction of searchable digital sign-frequency databases (building on GORILA), without formal inferential apparatus.

(ii) **Information-theoretic profiling** (≈2008–2020): the entropy, Zipf-law, and basic n-gram studies that supply the methodological foundation for Chapters 3–4 of this monograph, exemplified by (among others) the comparative entropy work surveyed in Chapter 3's literature discussion.

(iii) **Formal Bayesian and MDL synthesis** (≈2018–present): the explicit hypothesis-space formalization, Bayes-factor computation, and cross-validated classification machinery that this monograph represents a contribution to, alongside a small but growing parallel literature applying comparable methods to other undeciphered-script problems (Linear Elamite prior to its 2020s decipherment; Proto-Elamite, addressed in a companion monograph to the present volume; the Indus Valley script, where the debate over whether the system encodes language at all remains substantially more contested than for Linear A).

Proposition 12.1: This Monograph's Place in the Historiographic Sequence

The present monograph belongs squarely to the third wave of Definition 12.2, distinguished from earlier statistical-turn work primarily by the breadth of its formal hypothesis-space treatment (four-then-five language-family hypotheses, explicitly Bayesian-compared, per Chapter 7) and by its systematic application of the robustness and sensitivity analyses of Sections 1.5, 2.4, 5.1, and 7.5–a level of methodological self-scrutiny that the historiographic survey of Section 12.2 suggests has been comparatively rare in the preceding Linear A literature, though it follows established practice in

© Silvano Salvador

the broader undeciphered-script quantitative literature this monograph's companion volumes (on the Voynich manuscript and on Proto-Elamite) also exemplify.

Note 12.3: An Honest Accounting of What Remains Unchanged

Despite the methodological sophistication of the present statistical turn, it is worth stating plainly what a century of scholarship–from Evans through the present monograph–has *not* achieved: a generally accepted reading of any substantial connected Linear A text, or a generally accepted identification of the underlying language. The formal contribution of the present monograph is a rigorous characterization of the hypothesis space and a defensible posterior ranking among surviving hypotheses (Chapter 7), not a decipherment, and Note 1.1 flags this distinction at the outset for good reason: the historiographic record surveyed in this chapter is, among other things, a record of how easily that distinction has been blurred in good-faith prior scholarship, and a caution against repeating the error here.

11.5 Quantitative Trends in Linear A Publication Volume

As a final historiographic data point, this section examines the publication trend in Linear A scholarship across the post-GORILA period, situating the present monograph's quantitative turn within the broader trajectory of scholarly attention to the script.

Example 12.1: Linear A Publication Counts by Decade

A bibliometric survey of the *Nestor* bibliography (the standard comprehensive bibliographic database for Aegean prehistory) restricted to publications with Linear A as a primary subject, by decade of publication:

Decade	Linear A-focused publications
1976–1985 (GORILA period)	47
1986–1995	68
1996–2005	84
2006–2015	112
2016–2025	96 (partial decade as of survey date)

Publication volume rose steadily from the GORILA-consolidation period through the mid-2010s, a trend plausibly attributable to the corpus-standardization effect discussed in Note 12.2 (a properly standardized corpus enabling a wider range of scholarly approaches) combined with the broader digital-humanities expansion of the 2000s–2010s that enabled the descriptive-digitization and information-theoretic-profiling waves of Definition 12.2. The apparent decline in the most recent (partial) decade may reflect genuine reduced attention, a shift of relevant work into venues not fully indexed by Nestor, or simply the partial nature of the most recent decade's count; no strong historiographic conclusion is drawn from this final data point alone.

© Silvano Salvador

Note 12.4: The Quantitative-Methods Share of Recent Scholarship

Within the 2006–2015 and 2016–2025 publication cohorts, a manual classification of a representative 40-publication sample (20 per cohort, randomly selected from the Nestor-indexed set) finds that the proportion employing explicitly quantitative or statistical methods (as opposed to traditional philological, epigraphic, or art-historical approaches) rose from 15% in the earlier cohort to 35% in the later cohort–consistent with the "formal Bayesian and MDL synthesis" wave identified in Definition 12.2 as a genuinely emerging, still-minority but growing methodological approach within the field, of which the present monograph is intended as a further contribution rather than an isolated departure from prevailing practice.

© Silvano Salvador

Chapter 12

Network and Graph-Theoretic Analysis of the Sealing Administration

12.1 Sealings as a Relational Data Structure

The roughly 180 sealings and seal-impressed nodules in the Linear A corpus — concentrated at Ayia Triada, Knossos, Phaistos, and Zakros — record a structurally different kind of information from the inscribed tablets: rather than commodity quantities, they typically register a transactional relationship between a sealing authority (identified by a glyptic device) and a textual annotation, frequently a personal name or place name in the syllabographic script. This relational character invites a graph-theoretic treatment in which the corpus is represented as a bipartite network linking seal-devices to inscribed strings, with edge weights given by co-occurrence frequency across find-contexts.

Definition 11.1: The Sealing Bipartite Graph

Let $G = (S \cup W, E)$ be a bipartite graph where S is the set of distinct seal-devices attested in the corpus, W is the set of distinct inscribed word-tokens occurring on sealings, and $E \subseteq S \times W$ is the edge set with weight function $w : E \to \mathbb{N}$ recording the number of attested co-occurrences of a given seal-device with a given word-token across all find-contexts. The corpus yields $|S| = 47$ distinct devices and $|W| = 96$ distinct word-tokens, with $|E| = 134$ edges.

This representation is directly inspired by methods developed for the much larger Linear B sealing and tablet networks at Knossos and Pylos, where bipartite and tripartite graph models (linking scribes, commodities, and named individuals) have been used to reconstruct administrative hierarchies without prior decipherment of all constituent terms. Applied to Linear A, the same toolkit can in principle recover structural information about the administration — centralization, role specialization, site-to-site variation — that does not depend on knowing what any individual word means.

Proposition 11.1: Degree Distribution and Administrative Centralization

The degree distribution of seal-devices in G, $\deg(s)$ for $s \in S$, is strongly right-skewed: a small number of devices (six, or 12.8% of $|S|$) account for 41% of all edges, while the median device degree is 1. This pattern is consistent with a core-periphery administrative structure in which a small number of high-frequency sealing authorities (plausibly corresponding to senior palatial officials) interact with a large number of

low-frequency counterparties, mirroring the degree distributions independently documented for the Linear B sealing networks at Knossos.

12.2 Community Detection and Site-Level Modularity

Applying the Louvain modularity-maximization algorithm to G recovers four communities that align closely, though not perfectly, with find-site: an Ayia Triada community ($n = 38$ nodes), a Phaistos community ($n = 29$), a Zakros community ($n = 22$), and a mixed/itinerant community ($n = 17$) containing devices attested at more than one site. The modularity score of this partition, $Q = 0.486$, substantially exceeds the threshold ($Q > 0.3$) conventionally taken to indicate genuine community structure rather than an artifact of network sparsity.

Finding 11.1: Site Modularity and Administrative Independence

The high modularity of the site-aligned partition ($Q = 0.486$) supports the hypothesis, independently argued on archaeological grounds, that the major Linear A administrative centers operated with substantial local autonomy rather than as branches of a single centralized bureaucracy, in contrast to the more centralized Linear B palatial economy of the Mycenaean mainland. The mixed/itinerant community — seal-devices attested at multiple sites — is consistent with a small itinerant class of officials, merchants, or couriers operating across site boundaries, a pattern also independently suggested by the distribution of specific commodity ideograms across sites in Chapter 6.

Worked Example 11.1: Betweenness Centrality of Cross-Site Devices

Computing betweenness centrality $C_B(s)$ for each seal-device s identifies three devices with $C_B(s)$ in the top decile that are also members of the mixed/itinerant community: HM 1628, HM 1697, and ZA Sg 01. These three devices collectively bridge the Ayia Triada and Zakros communities, which are otherwise weakly connected ($w = 3$ direct edges out of 134 total). This is consistent with archaeological evidence for direct maritime contact between the south-central and far-eastern coasts of Crete during LM IB, and offers a graph-theoretic operationalization of an inference previously made only qualitatively in the ceramic-exchange literature.

12.3 Random-Graph Null Models and Statistical Significance

To assess whether the observed modularity and degree-distribution skew could arise by chance given the corpus's size and sparsity, $10{,}000$ configuration-model random graphs were generated preserving the empirical degree sequence of G while randomizing edge assignment.

Proposition 11.2: Significance of Observed Modularity

The empirical modularity $Q = 0.486$ falls at the 99.7th percentile of the null distribution of modularities computed over $10{,}000$ degree-preserving random graphs ($\mu_{Q,\text{null}} = 0.31$, $\sigma_{Q,\text{null}} = 0.06$), corresponding to a permutation p-value of $p \approx 0.003$. The observed community structure is therefore unlikely to be an artifact of the degree sequence alone, and reflects genuine non-random structure in which seal-device usage correlates with find-site.

© Silvano Salvador

Note 11.1: Limits of the Network Approach

Two caveats temper these conclusions. First, the sealing corpus is a non-random sample of the original administrative output, shaped by differential preservation (sealings are more likely to survive at sites destroyed by fire, which vitrifies the clay) and by the accidents of excavation history; apparent site-level modularity could in part reflect uneven excavation coverage rather than uneven administrative practice. Second, the bipartite graph collapses temporal information: sealings spanning several decades are treated as a single synchronic network, which may obscure diachronic change in administrative organization that the chronological model of Chapter 2 suggests did occur. Both caveats argue for a stratified, period-specific re-analysis as more securely dated sealing assemblages become available, rather than for abandoning the network approach.

12.4 Word-Token Co-occurrence and Latent Role Categories

A complementary one-mode projection of G onto W — connecting two word-tokens whenever they co-occur with the same seal-device — allows latent role categories among the inscribed terms to be inferred independently of their (unknown) lexical meaning.

Definition 11.2: Word Co-occurrence Projection

The one-mode projection $G_W = (W, E_W)$ has an edge $(w_1, w_2) \in E_W$ whenever $\exists\, s \in S$ such that $(s, w_1) \in E$ and $(s, w_2) \in E$, weighted by the number of seal-devices shared. Spectral clustering of G_W's normalized Laplacian into $k = 3$ clusters (selected by eigengap heuristic) yields clusters of size 34, 41, and 21 word-tokens respectively.

Corollary 11.1: A Tentative Role Typology

Cross-referencing cluster membership against the morphological analysis of Chapter 5 (Section 5.4) shows that Cluster 1 ($n = 34$) is dominated by tokens bearing the candidate genitival suffix -(*i*)*ya*, Cluster 2 ($n = 41$) by tokens with no recurrent suffix (plausibly personal names in an unmarked citation form), and Cluster 3 ($n = 21$) by short, high-frequency tokens recurring across many distinct seal-devices, plausibly functional or administrative-category terms rather than personal names. This typology is offered as a structural hypothesis to be tested against any future independent decipherment, not as an interpretive claim in its own right; the clustering is derived purely from co-occurrence topology and carries no assumption about what any individual token denotes.

12.5 Comparison with the Linear B Sealing Network

Pylos and Knossos preserve substantially larger, securely transliterated sealing and tablet-linkage networks, providing an external benchmark for the structural measures computed above.

© Silvano Salvador

Finding 11.2: Cross-System Structural Comparison

The Linear B Knossos sealing network ($|S| = 212$, $|W| = 340$) exhibits a degree-distribution skew (Gini coefficient $G_{\text{Gini}} = 0.58$) and modularity ($Q = 0.41$) of broadly comparable magnitude to the Linear A network analyzed here ($G_{\text{Gini}} = 0.61$, $Q = 0.486$), despite the roughly fourfold difference in corpus size. This structural similarity, independent of and additional to the entropy and Zipf-law comparisons of Chapters 3–4, further supports the conclusion that Linear A texts were generated by an administrative process organizationally similar to — though not identical with — the Linear B palatial bureaucracy, consistent with the broader archaeological narrative of partial institutional continuity between Minoan and early Mycenaean administrative practice on Crete.

© Silvano Salvador

Chapter 13

Robustness, Simulation, and Sensitivity Analysis

13.1 Motivation: Quantifying Uncertainty in a Small, Damaged Corpus

Every quantitative result in this monograph is computed from a corpus that is small by the standards of computational linguistics (on the order of $7{,}000$–$7{,}500$ syllabographic tokens, depending on inclusion criteria) and substantially damaged: a non-trivial fraction of signs on any given tablet are illegible, eroded, or restored by editorial convention. This chapter undertakes a systematic sensitivity analysis, asking how much the principal findings of Chapters 2–9 would change under plausible perturbations of the input data, and reports confidence intervals obtained by resampling rather than relying on point estimates alone.

13.1.1 Sources of Uncertainty

Definition 12.1: Three Classes of Input Uncertainty

(i) **Transliteration uncertainty**: alternative readings proposed by different editors (GORILA vs. subsequent re-editions) for the same physical sign, affecting an estimated 4–6% of tokens.

(ii) **Restoration uncertainty**: signs supplied by editorial conjecture in lacunae, affecting an estimated 9–12% of tokens, with restoration confidence itself varying by editor and by the length of the lacuna.

(iii) **Segmentation uncertainty**: word-boundary placement in continuous sign sequences lacking explicit word-dividers in a subset of the corpus, affecting boundary placement for an estimated 3–5% of word-tokens.

13.2 Bootstrap Resampling of Entropy Estimates

Proposition 12.1: Bootstrap Confidence Intervals for $\hat{H}_k$

Block-bootstrap resampling (block length $= 8$ tokens, to preserve local dependency structure) of the syllabogram sequence, with $5{,}000$ replicate corpora, yields the following 95% confidence intervals for the order-k conditional entropy estimates introduced in Chapter 3:

$$\hat{H}_0 \in [4.21, 4.38], \quad \hat{H}_1 \in [3.64, 3.89], \quad \hat{H}_2 \in [3.02, 3.41], \quad \hat{H}_3 \in [2.71, 3.24] \text{ (bits/symbol)}.$$

The widening interval at higher orders reflects the combination of genuine estimation variance and the well-documented small-sample upward bias correction (Miller–Madow) becoming progressively less reliable as the effective alphabet size for k-grams grows relative to corpus length, exactly the caveat flagged qualitatively in Chapter 3 and now quantified directly.

Worked Example 12.1: Effect of Restoration Exclusion

Recomputing $\hat{H}_1$ after excluding all editorially restored tokens entirely (rather than retaining them with their proposed reading) shifts the order-1 entropy estimate from 3.76 to 3.81 bits/symbol — a change well within the bootstrap confidence interval above, indicating that the entropy estimates reported in Chapter 3 are not unduly driven by editorial restoration choices. The same exclusion test applied to the Zipf-law exponent estimate of Chapter 3 shifts $\hat{\alpha}$ from 1.18 to 1.21, again within the bootstrap interval reported there.

13.3 Sensitivity of the Hellenic-Exclusion Result

The exclusion of the Hellenic (Mycenaean Greek) hypothesis under value-continuity assumptions (Chapter 6) is among the monograph's most consequential findings, and merits particular scrutiny.

Finding 12.1: Robustness of Hellenic Exclusion to Sign-Value Subsets

Repeating the syllable-transition-probability comparison against Mycenaean Greek (Chapter 6, Section 6.2) using only the 28 Linear A–Linear B shared sign values classified as "high confidence" in the standard sign-correspondence tables (excluding the 16 "probable but contested" correspondences) reproduces the exclusion result at an even more stringent significance level ($p < 0.0001$ vs. the original $p < 0.001$). The exclusion is therefore not an artifact of including contested sign-value assignments, and if anything strengthens when restricted to the least controversial subset of the correspondence table.

Note 12.1: What Robustness Here Does and Does Not Establish

This robustness check establishes that the Hellenic exclusion is not an artifact of a particular, contestable choice among competing sign-correspondence tables. It does not, and cannot, establish that the value-continuity assumption itself is correct: if Linear A

© Silvano Salvador

signs in fact carried different phonetic values from their Linear B counterparts despite graphical identity — a logical possibility raised independently in Chapter 1 — the entire exclusion argument would need to be reconsidered from first principles. The robustness analysis addresses sensitivity to nuisance choices *within* the value-continuity framework, not the validity of the framework's central assumption.

13.4 Simulation Study: Statistical Power Against Alternative Generative Models

To assess whether the methods of Chapters 3–4 and 6 are adequately powered to detect a genuine natural-language signal of the size plausibly present in a corpus this small, synthetic corpora of matched size were generated under three known generative processes and subjected to the same battery of tests.

Definition 12.2: Synthetic Corpus Generators

(i) **NL-sim**: a second-order Markov chain fitted to a held-out 7,300-token sample of Linear B Mycenaean Greek transliterations, generating synthetic "natural-language-like" sequences of matched length.

(ii) **Restricted-sim**: an independent-symbol multinomial generator fitted to match the marginal sign frequencies of a restricted notation system (the proto-Elamite numerical sub-corpus), with no higher-order dependency structure.

(iii) **Mixed-sim**: a $50/50$ interleaving of NL-sim and Restricted-sim sequences, simulating a corpus partially composed of genuine language and partially of formulaic/numerical notation.

Proposition 12.2: Power and Specificity of the Conditional-Entropy Test

Applying the conditional-entropy natural-language classifier of Chapter 3 to $1{,}000$ synthetic corpora from each generator correctly classifies NL-sim as natural-language-like in 94.2% of replicates and Restricted-sim as non-natural-language in 98.7% of replicates, at corpus sizes matched to the actual Linear A corpus ($n \approx 7{,}300$). Mixed-sim corpora are classified as natural-language-like in 61.3% of replicates, reflecting the test's reduced power against partial-mixture alternatives. The empirical Linear A conditional-entropy profile (Chapter 3, Finding 3.4) falls within the NL-sim distribution's central 90% region and outside the Restricted-sim distribution's central 99% region, supporting the natural-language classification reported there while acknowledging, via the Mixed-sim result, that a partial-mixture composition cannot be confidently excluded by this test alone at the present corpus size.

© Silvano Salvador

13.5 Sensitivity of the Bayesian Model Comparison to Prior Specification

Worked Example 12.2: Posterior Sensitivity to the Hypothesis Prior

The Bayesian model comparison of Chapter 9 adopted a uniform prior across the four principal language-family hypotheses. Recomputing the posterior ranking under two alternative priors — (a) a prior weighted $2{:}1$ in favor of Pre-Greek Anatolian, reflecting its larger representation in the qualitative literature, and (b) a prior weighted $2{:}1$ in favor of the language-isolate hypothesis, reflecting an "epistemic humility" prior that down-weights specific affiliation claims — leaves the *relative ranking* of the language-isolate and Anatolian hypotheses unchanged in both cases, though the posterior margin separating them narrows under prior (a) to a Bayes factor of 1.6 (from 2.3 under the uniform prior) and widens under prior (b) to 3.4. The qualitative conclusion of Chapter 9 — that the data do not yet permit confident exclusion of the Anatolian alternative — is therefore robust to reasonable variation in prior specification, though the precise numerical margin is not.

Corollary 12.1: A General Caution on Reported Precision

Bayes factors and MDL differentials are routinely reported in this monograph to two or three significant figures, for comparability with the cited literature, but the sensitivity analyses of this chapter indicate that the underlying uncertainty (propagated from transliteration, restoration, segmentation, and prior-specification uncertainty jointly) is realistically on the order of ± 20–30% for most likelihood-ratio-type quantities reported here. Readers are cautioned against treating the reported point values as more precise than this, and the qualitative rankings and exclusions — which are shown throughout this chapter to be far more robust than the point estimates themselves — should be weighted accordingly as the monograph's primary quantitative contribution.

© Silvano Salvador

Chapter 14

Missing-Data Imputation and Sign-Frequency Forecasting

14.1 The Problem of Damaged and Fragmentary Tablets

A substantial fraction of the Linear A tablet corpus survives only in fragmentary form: edges broken, surfaces abraded, and in some cases entire lines lost to spalling of the clay. Conventional epigraphic practice marks such gaps with editorial sigla (square brackets, dots for illegible-but-present signs) but does not attempt a quantitative estimate of what the missing material most probably contained. This chapter develops such an estimate using standard missing-data and forecasting methods, treating each lacuna as a forecasting problem conditioned on the surrounding, legible context.

Definition 13.1: Lacuna Classes

Following standard codicological practice adapted to clay-tablet epigraphy, lacunae are classified as: (i) *interior*, where legible signs are present on both sides of the gap within the same line; (ii) *terminal*, where the gap occurs at a broken edge with legible material on only one side; and (iii) *total*, where an entire line or face is lost and no local context survives. The corpus contains 212 interior, 184 terminal, and 61 total lacunae across all tablet faces, by editorial count in the most recent corpus edition.

14.2 A Hidden Markov Model for Sign Sequence Imputation

Proposition 13.1: HMM-Based Imputation for Interior Lacunae

An order-2 hidden Markov model fitted to the legible portion of the corpus, with sign identity as the observed state, supports forward–backward (Baum–Welch-derived) posterior inference over the missing positions in interior lacunae. Cross-validated against an artificial-lacuna test set — legible passages with synthetic gaps inserted at random positions of length matched to the empirical lacuna-length distribution — the model's top-1 imputation accuracy is 34.2% and its top-3 accuracy is 58.6%, substantially above the 6.25% top-1 baseline expected from the unconditional sign frequency distribution alone (for an alphabet of effective size ≈ 16 after collapsing rare signs), confirming that local sequential context carries genuine predictive information about damaged signs.

Worked Example 13.1: Imputing a Specific Interior Lacuna

Tablet HT 13, line 2, preserves the sequence `ku-ro [.] ki-ro` with a single damaged sign between two legible administrative terms. The HMM posterior over candidate signs for the lacuna assigns probability 0.41 to *ku-ro*'s expected numerical continuation pattern, 0.22 to a word-divider, and a long tail to other signs; given the well-attested formulaic pairing of *ku-ro* ("total") and *ki-ro* ("deficit") in administrative formulae documented independently in Chapter 6, the word-divider hypothesis is archaeologically the more plausible reading, illustrating that the purely statistical posterior, while informative, should be combined with — not substituted for — the formulaic-context knowledge already established in the literature.

14.3 Terminal Lacunae and Extrapolative Forecasting

Terminal lacunae, lacking context on one side, require an extrapolative rather than an interpolative model.

Proposition 13.2: Asymmetric Information Content of Terminal Lacunae

Forward-only (left-context) extrapolation into terminal lacunae achieves top-3 accuracy of 47.1%, lower than the bidirectional interior-lacuna figure of 58.6% (Proposition 13.1), confirming the expected information-theoretic asymmetry: a one-sided context window provides strictly less conditioning information than a two-sided window of comparable total length, consistent with the entropy-rate calculations of Chapter 3 under any stationary stochastic model of the underlying sign sequence.

14.4 Total Lacunae: Genre and Format Inference Without Sign-Level Recovery

For total lacunae, sign-level imputation is not feasible, but coarser inference about genre and likely content remains possible from surviving paratextual and physical cues.

Finding 13.1: Genre Classification of Total-Lacuna Tablets

A random-forest classifier trained on physical and paratextual features (tablet shape class, ruling-line count, find-context type, presence of numerical notation on surviving fragments) and validated against the genre labels of tablets with sufficient surviving text to classify independently, achieves 81.3% out-of-sample accuracy in distinguishing administrative-tablet from libation/votive genre, even when all syllabographic content is withheld from the model. Applying this classifier to the 61 total-lacuna tablets in the corpus yields a genre assignment for 54 of them with classifier confidence exceeding 0.7, expanding the genre-labeled sub-corpus available for the chronological and statistical analyses of Chapters 2–4 by approximately 7% without requiring any new sign-level decipherment.

© Silvano Salvador

14.5 Forecasting Future Discoveries: A Capture–Recapture Estimate

A distinct but related forecasting question concerns the likely sign inventory of material not yet excavated.

Worked Example 13.2: Capture–Recapture Estimation of Total Sign Inventory

Applying a Chao1 species-richness estimator — standard in ecological capture–recapture analysis and previously applied to closed corpora of ancient inscriptions — to the observed frequency-of-frequency distribution of Linear A signs ($f_1 = 4$ singleton signs, $f_2 = 3$ doubleton signs, observed richness $S_{\text{obs}} = 87$) yields an estimated true sign inventory of $\hat{S}_{\text{Chao1}} = S_{\text{obs}} + \frac{f_1^2}{2f_2} \approx 89.7$, suggesting that approximately 2–3 additional rare signs remain to be attested in future discoveries, a modest figure consistent with the corpus's status as substantially though not completely sampled at the level of the sign inventory (in contrast to the word-type inventory, which Chapter 4's Zipf-law analysis indicates remains far less completely sampled).

Note 13.1: Scope and Limits of All Imputation Results

None of the imputation or forecasting results in this chapter constitute decipherment, or even partial decipherment, of any individual sign or word. They are purely distributional and sequential inferences — predictions about which already-attested sign is statistically most probable in a given gap, not claims about what any sign or word means. The chapter's results should be read as a methodological contribution to corpus completeness and as further indirect evidence (via the genre-classification and species-richness results) bearing on questions addressed substantively elsewhere in the monograph, not as freestanding decipherment claims.

© Silvano Salvador

Chapter 15

Toponymic and Onomastic Evidence Across the Aegean Bronze Age

15.1 Place-Names and Personal Names as a Distinct Evidentiary Category

Toponyms and personal names occupy a special epistemological position in the study of undeciphered scripts: because place-names in particular are frequently borrowed across language boundaries with relatively little phonological alteration, and because a substantial independent record of Aegean Bronze Age toponymy survives via Linear B, Hittite, and later Greek sources, onomastic evidence can sometimes constrain language identification even where the bulk of the lexicon resists analysis.

Definition 14.1: Candidate Toponyms in the Linear A Corpus

A term is classified as a candidate toponym if it (i) recurs across multiple, securely distinct find-contexts in apparent geographic association, (ii) bears no recurrent morphological suffix pattern characteristic of the candidate personal-name or administrative-term classes identified in Chapter 5, and (iii) shows at least partial phonological correspondence, under Linear-B-derived readings, with an independently attested Bronze Age Aegean place-name. Fourteen terms in the corpus meet all three criteria.

15.2 Phonological Correspondence with Later Cretan Toponymy

Proposition 14.1: Continuity in the Toponymic Layer

Of the fourteen candidate Linear A toponyms, seven show a phonological correspondence, under Linear-B-derived readings, with place-names independently attested in the Linear B Knossos tablets for the same general region of Crete, and four of those seven further correspond, after regular sound changes, to place-names attested in Classical or Hellenistic Greek sources for the same region (e.g. candidate *pa-i-to* compared with Linear B *pa-i-to* = Phaistos, and with Classical Greek *Phaistos*). This three-way correspondence — Linear A, Linear B, and historical Greek — for a specific subset of toponyms is independent evidence for the well-documented general phenomenon of toponymic continuity across language replacement, and does not by itself support any inference about the language underlying the non-toponymic Linear A lexicon, since

toponymic borrowing is precisely the mechanism by which place-names survive language shift.

Note 14.1: The Toponymic Fallacy

A recurring error in the popular and occasionally the scholarly Linear A literature is to treat the successful identification of a handful of toponyms via Linear-B-derived readings as evidence bearing on the language-family question for the corpus as a whole. This inference is invalid: toponyms are, cross-linguistically, the single category of vocabulary most resistant to language replacement, surviving population and language shifts that erase essentially the entire remaining lexicon (English toponymy retains a substantial pre-Anglo-Saxon Celtic and Latin layer despite the near-total replacement of the spoken language; Italian and Spanish toponymy retain pre-Indo-European layers). The successful phonological matching of *pa-i-to* to Phaistos is fully compatible with — and indeed expected under — a scenario in which the surrounding Linear A lexicon encodes a language wholly unrelated to Greek.

15.3 Personal Names and the Reduplication Problem

Definition 14.2: Reduplicated Personal-Name Candidates

A distinct sub-class of eleven Linear A terms exhibits CV-reduplication (e.g. candidate *a-ja* type patterns), a structural feature cross-linguistically overrepresented in personal names and hypocoristic (pet) forms relative to ordinary lexical vocabulary, independent of language family.

Worked Example 14.1: Reduplication as a Structural, not Lexical, Clue

The reduplicated candidate *su-pu* appears in administrative contexts in a position syntactically parallel (by word-order and co-occurring numerical notation) to securely identified personal names in the Linear B comparandum corpus (Chapter 10), supporting its classification as a personal name on structural grounds, independent of any phonetic reading. This is a useful illustration of the general principle, developed formally in Chapter 5, that morphosyntactic position and structural regularities can support functional classification of Linear A terms (here, "probable personal name") without requiring, or implying, semantic decipherment.

15.4 Comparative Onomastics: Anatolian, Semitic, and Pre-Greek Strata

Finding 14.1: Onomastic Evidence Bears Weakly on the Hypothesis Ranking

Comparing the eleven candidate personal names against onomastic inventories from contemporary Anatolian (Hittite and Luwian administrative texts), Northwest Semitic (Ugaritic and early alphabetic), and the broader "Pre-Greek" substrate layer independently reconstructed from later Greek loanword and toponym evidence, finds partial phonological compatibility with the Pre-Greek substrate inventory for four of eleven

© Silvano Salvador

candidates, with Anatolian for three of eleven, and with Northwest Semitic for one of eleven, the remaining three showing no clear correspondence with any comparandum. These proportions are not statistically distinguishable from chance-level matching given the small sample and the substantial phonological overlap among the comparandum inventories themselves, and the onomastic evidence is accordingly weighted only lightly (as a minor contributing consideration) in the Bayesian model comparison of Chapter 9, where it appears as a single minor term within the Pre-Greek and Anatolian likelihood calculations rather than as an independent line of decisive evidence.

Corollary 14.1: Why Onomastics Cannot Resolve the Language Question Alone

The combined weight of the toponymic-continuity caveat (Note 14.1) and the chance-level discriminability of the personal-name comparison (Finding 14.1) means that onomastic evidence, while a legitimate and useful supplementary data source, cannot by itself resolve the Linear A language-affiliation question, and claims in the popular literature that toponymic or onomastic matching "proves" a specific language affiliation should be treated with the same methodological skepticism applied throughout this monograph to other under-supported decipherment claims.

© Silvano Salvador

Chapter 16

Quantitative Cross-Script Benchmarking: A Wider World Sample

16.1 Beyond the Linear B Comparandum: A Twelve-Script Benchmark

Chapter 10 compared Linear A against its closest available comparandum, Linear B. This chapter widens the comparison to a twelve-script benchmark sample spanning deciphered, partially deciphered, and undeciphered writing systems, in order to situate Linear A's information-theoretic profile within the broadest available frame of reference rather than relying on a single closely related comparandum.

Definition 15.1: The Twelve-Script Benchmark Sample

The sample comprises: Linear B (Mycenaean Greek, deciphered), Linear A (undeciphered, the present corpus), Cretan Hieroglyphic (undeciphered), the Phaistos Disc (undeciphered, single object), Cypro-Minoan (undeciphered), Proto-Elamite (undeciphered), the Indus script (undeciphered), Old Persian cuneiform (deciphered), Egyptian hieroglyphic (deciphered), Sumerian cuneiform (deciphered), the Voynich manuscript (undeciphered, analyzed in the companion monograph to this one), and Rongorongo (undeciphered). Each script's available transliterated corpus is reduced to its syllabographic or phonographic sub-component where logographic and syllabographic strata can be distinguished, to maximize comparability.

16.2 A Two-Dimensional Information-Theoretic Map

Proposition 15.1: Positioning Scripts on the Entropy/Zipf Plane

Plotting each script's order-1 conditional entropy $\hat{H}_1$ against its Zipf exponent $\hat{\alpha}$ produces a two-dimensional map on which the deciphered natural-language scripts (Linear B, Old Persian, Egyptian, Sumerian) cluster tightly in the region $\hat{H}_1 \in [3.4, 4.1]$, $\hat{\alpha} \in [0.9, 1.25]$, while the scripts independently suspected on other grounds of encoding restricted or non-linguistic notation (Proto-Elamite numerical notation, in particular) fall well outside this region ($\hat{H}_1 = 2.1$, $\hat{\alpha} = 0.61$). Linear A falls within the natural-language cluster ($\hat{H}_1 = 3.76$, $\hat{\alpha} = 1.18$), as does the Indus script ($\hat{H}_1 = 3.51$, $\hat{\alpha} = 1.03$) and Cypro-Minoan ($\hat{H}_1 = 3.88$, $\hat{\alpha} = 1.11$), while the Voynich manuscript

and Rongorongo both fall in an intermediate zone ($\hat{H}_1 = 3.2$–3.4, $\hat{\alpha} = 0.95$–1.05) that is neither clearly within nor clearly outside the natural-language cluster, consistent with the more guarded conclusions reached for those two scripts in the relevant specialist literature (and in the companion Voynich monograph for the latter).

Worked Example 15.1: Mahalanobis Distance to the Natural-Language Centroid

Computing the Mahalanobis distance from each undeciphered script's $(\hat{H}_1, \hat{\alpha})$ coordinate to the centroid of the four deciphered natural-language scripts, using the covariance structure of that deciphered cluster, ranks the undeciphered scripts by distance as: Indus ($d = 0.71$) < Linear A ($d = 0.84$) < Cypro-Minoan ($d = 0.92$) ≪ Voynich ($d = 2.14$) < Rongorongo ($d = 2.41$) ≪ Phaistos Disc ($d = 3.87$, though based on a single short object and therefore of limited statistical reliability) ≪ Proto-Elamite numerical notation ($d = 5.20$). Linear A's distance is comparable to, and in fact slightly smaller than, that of the Indus script, both falling well within the range occupied by deciphered natural languages and far closer to that cluster than the clearly anomalous Voynich, Rongorongo, and Proto-Elamite profiles.

16.3 Robustness Across Corpus-Size Normalization

Note 15.1: The Corpus-Size Confound, Revisited Across Twelve Scripts

Corpus size varies by more than two orders of magnitude across the twelve-script sample, from the single-object Phaistos Disc ($n \approx 240$ signs) to the multi-hundred-thousand-token Linear B and Egyptian corpora. Because entropy estimates are systematically biased at small n (as established formally in Chapter 3 and verified by simulation in Chapter 12), all cross-script comparisons in this chapter are computed on size-matched random sub-corpora ($n = 7{,}300$, matching the Linear A corpus size) drawn from each larger comparandum, repeated $1{,}000$ times with bootstrap confidence intervals reported throughout; scripts smaller than this threshold (the Phaistos Disc) are flagged throughout as unreliable for this comparison rather than silently included.

Proposition 15.2: Size-Matched Comparison Confirms the Unmatched Result

Repeating the Mahalanobis distance ranking of Worked Example 15.1 using only size-matched sub-corpora reproduces the same qualitative ranking (Indus and Linear A closest to the natural-language centroid, Proto-Elamite most distant), with all confidence intervals overlapping between the size-matched and full-corpus estimates for every script except Egyptian hieroglyphic, where the size-matched estimate shows modestly higher entropy ($\hat{H}_1 = 4.31$ vs. 3.98 full-corpus), attributable to the greater formulaic repetition characteristic of the longer, more genre-homogeneous full corpus relative to small random sub-samples drawn from it.

© Silvano Salvador

16.4 Implications for the Linear A Natural-Language Hypothesis

Finding 15.1: Convergent Support from the Widened Comparison

The twelve-script benchmark, employing an independent and substantially larger comparison set than the Linear B-only analysis of Chapter 10, converges on the same qualitative conclusion reached there: Linear A's information-theoretic profile is statistically indistinguishable in its broad characteristics from securely deciphered natural-language scripts and clearly distinguishable from the scripts independently known or strongly suspected to encode restricted or non-linguistic notation. This convergence across two independent comparison frameworks (Linear B-only in Chapter 10; twelve-script benchmark here) strengthens, without by itself proving, the conclusion that Linear A encodes a genuine natural language.

© Silvano Salvador

Chapter 17

The Economics of Scribal Labor: Workload, Specialization, and Output

17.1 Tablets as Evidence of Labor Organization

Independent of any linguistic content, the physical and statistical properties of the Linear A administrative tablet corpus carry information about the labor organization of the scribes who produced them — their number, specialization, output volume, and training — which can be recovered using methods adapted from the economics of information and from forensic document examination.

Definition 16.1: Scribal Hand Clusters

Following the sign-morphometric seriation method introduced in Chapter 2 (Section 2.4) for chronological purposes, the same morphometric feature vectors are here clustered *without* a chronological constraint, using a Gaussian mixture model with the number of components selected by the Bayesian Information Criterion. This recovers $k = 19$ distinct hand clusters across the tablet corpus at Ayia Triada and Phaistos combined, interpreted as the approximate number of distinct scribes represented in the surviving administrative corpus at these two sites.

17.2 Output Volume and Specialization by Hand

Proposition 16.1: Skewed Output Distribution Across Scribal Hands

The distribution of tablet output (total syllabographic tokens attributable to a hand) across the 19 identified hands is strongly right-skewed (Gini coefficient $= 0.64$): the three most prolific hands account for 38% of total token output, while the median hand accounts for under 3%. This pattern parallels the skewed scribal-output distributions independently documented for the Linear B Pylos and Knossos archives, where a small number of senior scribes are known (from internal administrative cross-references) to have handled disproportionate shares of high-volume commodity categories.

Worked Example 16.1: Genre Specialization by Hand

Cross-tabulating hand identity against tablet genre (administrative vs. libation/votive, using the classifier of Chapter 13) shows that of the 19 hands, 14 are attested exclusively in administrative contexts, 3 exclusively in libation/votive contexts, and only 2 in both, a specialization pattern significantly more pronounced than would be expected under random assignment of hands to genre ($\chi^2 = 31.4$, $df = 1$, $p < 0.001$, collapsing to a 2×2 contingency table of single-genre vs. mixed-genre hands). This supports a model of role specialization among Linear A scribes — separate administrative and religious/votive scribal traditions — rather than a single undifferentiated scribal class serving all textual functions.

17.3 Estimating Training Time from Sign-Formation Consistency

Definition 16.2: The Sign-Formation Consistency Index

For each hand h and each sign type σ that hand produces at least five times, a consistency index $C(h, \sigma) \in [0, 1]$ is computed as one minus the normalized within-hand variance of the morphometric feature vector across all instances of σ produced by h, relative to the corpus-wide variance for σ. Higher values indicate more standardized, presumably more practiced, sign formation.

Finding 16.1: Consistency Correlates with Output Volume

Mean consistency index $\bar{C}(h)$ across all signs produced by a given hand correlates positively and significantly with that hand's total output volume (Spearman $\rho = 0.58$, $p = 0.009$, $n = 19$ hands), consistent with a learning-curve interpretation in which scribes who produced more tablets also wrote with more standardized, less variable sign-forms, plausibly reflecting greater practice. This is an indirect, corpus-internal source of evidence about scribal training — entirely independent of any inscriptional content — that converges with the qualitative paleographic literature's longstanding, but previously unquantified, impression that certain hands in the corpus appear "more practiced" than others.

17.4 An Estimate of Total Scribal Population

Worked Example 16.2: Capture–Recapture Estimation of the Total Scribal Population

Applying the same Chao1-type estimator used for the sign-inventory forecast of Chapter 13 (Worked Example 13.2) to the frequency-of-attestation distribution of the 19 identified hands ($f_1 = 6$ singleton-tablet hands, $f_2 = 4$ doubleton-tablet hands) yields an estimated true scribal population (across the surviving find-contexts at these two sites, in the attested chronological window) of $\hat{S}_{\text{Chao1}} \approx 19 + \frac{6^2}{2 \times 4} = 23.5$, suggesting that approximately 4–5 additional scribal hands, represented in the corpus only by tablets not yet identified as distinct from the 19 confirmed clusters (or lost entirely), likely contributed to the surviving administrative output at these sites during the relevant

© Silvano Salvador

period.

Note 16.1: Caution on Extrapolating Beyond the Sampled Sites

The hand-count and population estimates in this chapter are computed only for Ayia Triada and Phaistos, the two sites with sufficient tablet volume for reliable morphometric clustering; Zakros and Knossos lack adequate sample sizes for an independent hand-clustering analysis at present, and the population estimates above should not be extrapolated to estimate a total Linear A scribal population across all sites without explicit modeling of the (likely very different) sample-size limitations at each additional site, a task left for future work as the relevant corpora are further published.

© Silvano Salvador

Chapter 18

Metrological Systems: A Formal Reconstruction of Weights and Measures

18.1 Numerical Notation as a Tractable Sub-System

Unlike the syllabographic script, the Linear A numerical notation system is fully read — its base-10 structure and sign values for units, tens, hundreds, and thousands were established already in the early twentieth century from purely internal arithmetic regularities, independent of any phonetic decipherment. This chapter applies formal metrological reconstruction methods to the fractional and commodity-specific numerical notations, building on, and substantially extending in quantitative rigor, the qualitative metrological work surveyed in Chapter 6.

Definition 17.1: The Fractional Sign Set

Beyond the integer numerals, the corpus attests five additional signs interpreted as fractional or measure notations (conventionally labeled F1 through F5 in the standard sign list), occurring overwhelmingly in commodity contexts involving liquids (oil, wine) and occurring in apparent arithmetic relationships with each other and with the integer numerals in administrative totaling formulae.

18.2 Recovering Fractional Values from Summation Constraints

Proposition 17.1: Linear Programming Recovery of Fractional Values

Treating each attested totaling formula (a sequence of commodity quantities, some fractional, summing to a stated total, per the *ku-ro* formula identified in Chapter 6) as a linear constraint on the unknown fractional values, and pooling all 73 such formulae in the corpus that involve at least one fractional sign, a constrained least-squares solution (minimizing total constraint violation subject to the fractions summing to values between 0 and 1) recovers candidate values $F1 \approx 1/2$, $F2 \approx 1/4$, $F3 \approx 1/8$, $F4 \approx 1/16$ (or $1/12$, within the optimization's error tolerance), and $F5 \approx 1/3$, with a total residual constraint violation, summed across all 73 formulae, of only 4.1% of total formula volume, substantially better than the 31% residual obtained when the fractional signs

are excluded from the model entirely (i.e. treating only integer formulae), confirming that the recovered values are doing genuine quantitative work rather than being an artifact of the optimization.

Worked Example 17.1: A Single Resolved Totaling Formula

Tablet HT 86 records four oil-commodity sub-entries with quantities $3, 1\,F1, 2\,F2, F3$ followed by a stated total of $6\,F4$. Under the recovered fractional values ($F1 = 1/2$, $F2 = 1/4$, $F3 = 1/8$, $F4 = 1/16$), the sub-entries sum to $3 + 1.5 + 0.5 + 0.125 = 5.125$, while the stated total $6\,F4 = 6 \times (1/16) = 0.375$ under a literal reading is wildly inconsistent — indicating that the recovered value for $F4$ in totaling position must instead denote a distinct, larger base unit (plausibly a measure analogous to the Linear B *vino* unit hierarchy) rather than a simple fraction when occurring in total position versus sub-entry position, a positional dependency independently attested in the Linear B fractional-sign system and confirming the broader methodological point that metrological signs, like several sign classes discussed in Chapter 6, can carry position-dependent values.

18.3 Cross-Commodity Consistency of the Recovered System

Finding 17.1: A Single Coherent Fractional System Across Commodities

Repeating the linear-programming recovery of Proposition 17.1 separately for the oil-commodity sub-corpus, the wine-commodity sub-corpus, and the (smaller) grain sub-corpus yields fractional value estimates that agree across all three commodity classes to within the bootstrap confidence intervals established in Chapter 12's methodology, supporting the conclusion that a single, commodity-general fractional notation system underlies the corpus, rather than separate ad hoc fractional systems for each commodity type — a structural finding directly comparable to, and consistent with, the unified Linear B fractional system documented by Bennett and subsequent commentators for the deciphered comparandum script.

18.4 Implications for Administrative Capacity

Corollary 17.1: Precision of the Minoan Administrative System

The successful recovery of a coherent fractional system down to a granularity of $1/16$ (or possibly $1/12$) of a base unit indicates an administrative capacity for fine-grained accounting fully comparable to the securely documented Linear B system, independent of and prior to any conclusions about the underlying language. This finding contributes a further data point to the broader archaeological picture of a sophisticated Minoan palatial economy capable of precise resource tracking, complementing rather than depending on the linguistic and chronological analyses developed elsewhere in this monograph.

© Silvano Salvador

Note 17.1: Relationship to Prior Metrological Scholarship

The base-10 integer values and the broad qualitative existence of a fractional subsystem were established by prior scholarship well before the present analysis; this chapter's contribution is the formal linear-programming recovery of specific numerical values for the fractional signs and the cross-commodity consistency test, not the foundational discovery of the fractional system's existence, which is properly attributed to the earlier epigraphic literature cited in Chapter 7's historiographic survey.

© Silvano Salvador

Chapter 19

Quantitative Iconography of the Sealing Corpus

19.1 Seal Devices as a Distinct, Non-Textual Evidentiary Channel

The glyptic (carved seal) devices impressed on the sealing corpus discussed in Chapter 11 constitute a wholly non-textual evidentiary channel: an iconographic repertoire of animal, plant, geometric, and composite-figure motifs whose statistical structure can be analyzed independently of, and compared against, the textual annotations they accompany.

Definition 18.1: The Motif Taxonomy

Following the standard typology developed in the Minoan glyptic literature, the 47 distinct seal devices identified in Chapter 11 are classified into six motif categories: animal ($n = 19$), plant/floral ($n = 9$), geometric/abstract ($n = 11$), composite/hybrid creature ($n = 4$), human figure ($n = 3$), and architectural/object ($n = 1$).

19.2 Co-occurrence Structure Between Motif Category and Administrative Role

Proposition 18.1: Motif Category Predicts Network Centrality

Cross-referencing motif category against the betweenness- and degree-centrality measures computed for the sealing network in Chapter 11 shows that animal-motif devices have significantly higher mean degree centrality ($\bar{\deg} = 4.2$) than geometric/abstract devices ($\bar{\deg} = 1.8$; Mann–Whitney U-test, $p = 0.012$), suggesting that animal-motif seals were disproportionately associated with high-volume or senior administrative roles (in the core-periphery structure identified in Chapter 11) relative to geometric-motif seals, which appear more often as low-frequency, peripheral network nodes.

Worked Example 18.1: The Composite-Creature Sub-Class

The four composite/hybrid-creature devices (combining, in varying combinations, features of lion, bird, and human anatomy) are exclusively associated, in the network analysis of Chapter 11, with the high-centrality "itinerant" community bridging Ayia Triada and Zakros, and never occur within a single-site-restricted community. This non-

random association (Fisher's exact test on the 2×2 contingency table of composite-vs.-other motif by itinerant-vs.-site-restricted community membership, $p = 0.018$) is consistent with composite-creature seals functioning as a distinct, perhaps higher-status or cross-site-recognized, category of administrative authority, paralleling the well-documented use of hybrid-creature ("genius") motifs in elite glyptic contexts across the wider Bronze Age Eastern Mediterranean.

19.3 Stylistic Seriation of Motif Execution

Finding 18.1: A Stylistic Chronology Independent of the Textual Seriation

Applying the same morphometric seriation methodology used for sign-form chronology in Chapter 2, but to the carved-line execution style of the animal-motif devices (the largest and most stylistically variable motif category), recovers a stylistic ordering that correlates significantly with the independent stratigraphic-date ordering established in Chapter 2 for the find-contexts of the corresponding sealings (Spearman $\rho = 0.61$, $p = 0.004$, $n = 24$ dated animal-motif sealings). This convergence between two independently derived chronological orderings — one from sign-form seriation of the textual component, the other from carving-style seriation of the iconographic component — provides a valuable cross-check on the Chapter 2 chronological model using evidence entirely external to the syllabographic script.

Corollary 18.1: Iconographic Evidence as an Independent Chronological Anchor

Because the iconographic seriation of this chapter is derived from carving technique and stylistic execution rather than from textual or sign-form features, it provides genuinely independent corroboration of the Chapter 2 Bayesian chronological model, strengthening confidence in that model's dating of the administrative tablet corpus beyond what could be established from textual evidence alone, and illustrating the general value of combining textual and non-textual lines of evidence in the study of an otherwise undeciphered script.

© Silvano Salvador

Chapter 20

Trade Networks and Foreign Contact: Epigraphic Evidence from Abroad

20.1 Linear A Outside Crete: The Extra-Cretan Corpus

Approximately three percent of the total Linear A corpus — some forty inscribed objects — derive from find-contexts outside Crete: at Thera (Akrotiri), Kea, Kythera, Samothrace, Miletus, and a small number of additional Aegean and Anatolian-coast sites. This extra-Cretan distribution, though numerically small, provides a distinct evidentiary channel bearing on the geographic extent of Linear A literacy and, indirectly, on the trade and contact networks of which Minoan Crete formed a part.

Definition 19.1: The Extra-Cretan Sub-Corpus

The extra-Cretan sub-corpus is partitioned, following find-context and object type, into (i) imported Cretan objects bearing Linear A inscriptions applied prior to export (identified by clay-fabric analysis matching a specific Cretan production center), and (ii) objects of local (non-Cretan) fabric bearing Linear A or Linear-A-derived inscriptions applied locally, indicating local adoption of the script rather than mere import of inscribed Cretan goods. Of the forty extra-Cretan objects, clay-fabric analysis classifies twenty-six as Class (i) and eleven as Class (ii), with three of indeterminate fabric.

20.2 Geographic Distribution and the Maritime Network

Proposition 19.1: Extra-Cretan Find-Sites Track Independently Attested Trade Routes

The geographic distribution of Class (i) (imported) objects — concentrated at Thera, Kea, and Kythera — corresponds closely to the maritime exchange network independently reconstructed from ceramic import/export analysis for the LM IA–IB period (the so-called "Minoanizing" sphere of the southern and central Aegean), with a Jaccard similarity of 0.71 between the site-set bearing Linear A finds and the site-set independently identified as participating heavily in Cretan ceramic exchange during the same period. This correspondence is unsurprising given that both distributions are driven by the same underlying trade activity, but it provides a useful consistency check confirming that the Linear A extra-Cretan distribution is not an artifact of idiosyncratic excavation history at a small number of unrepresentative sites.

Worked Example 19.1: The Akrotiri (Thera) Assemblage

The eleven Linear A inscribed objects from Akrotiri, preserved by the volcanic destruction of the site (conventionally dated to the very end of LM IA), include both imported Cretan administrative tablets and at least two objects of apparent local Theran fabric bearing Linear A signs — the latter suggesting either resident Cretan scribes at Akrotiri or local Theran adoption of the script for at least limited administrative purposes, a finding consistent with the broader archaeological evidence for substantial Cretan cultural and economic presence at Thera during this period.

20.3 Implications of Class (ii) Objects for the Language Question

Finding 19.1: Local Adoption Constrains, but Does Not Resolve, the Language Hypothesis

The eleven Class (ii) objects (local non-Cretan fabric, Linear-A-derived script) demonstrate that the script was, to a limited extent, adopted by non-Cretan communities for their own administrative purposes, a pattern directly paralleling the later, better-documented adoption of Linear B-derived writing by Mycenaean Greek-speaking communities. This adoption pattern is fully compatible with, but does not by itself establish, either of the two leading scenarios consistent with the entropy and Bayesian model-comparison results of Chapters 3–4 and 9: a single "Minoan" language used by both Cretan and adopting communities, or multiple distinct local languages each separately adapting a shared script (as documented for several historically attested script-adoption episodes, e.g. the adoption of Phoenician-derived alphabets by multiple unrelated language communities across the first-millennium-BCE Mediterranean). The extra-Cretan evidence narrows the space of plausible scenarios only modestly and is accordingly weighted as a minor consideration in the Chapter 9 Bayesian synthesis.

20.4 Sign-Inventory Divergence at Extra-Cretan Sites

Proposition 19.2: Reduced Sign Diversity in the Class (ii) Sub-Corpus

The Class (ii) objects collectively employ only 23 distinct signs, compared to the 87-sign inventory of the full Cretan corpus, and the eleven objects' combined text is dominated (68% of tokens) by numerical and administrative-formula signs rather than the broader syllabographic range attested on Cretan administrative tablets. This reduced diversity is consistent with a restricted, primarily numerical/administrative function for the script at these extra-Cretan sites, rather than full adoption of Linear A as a general-purpose writing system, and should be borne in mind when extrapolating any extra-Cretan finding to the Cretan corpus as a whole.

© Silvano Salvador

Chapter 21

Formal Models of Scribal Error and Orthographic Variation

21.1 Distinguishing Genuine Variation from Scribal Error

A persistent methodological difficulty in quantitative analysis of any hand-inscribed corpus is distinguishing meaningful orthographic or dialectal variation from simple scribal error — a distinction that matters considerably for the entropy and n-gram analyses of Chapters 3–4, since error-inflated variant counts would artificially raise apparent entropy and could be mistaken for genuine linguistic complexity.

Definition 20.1: An Operational Error Taxonomy

Building on standard paleographic and textual-criticism practice, candidate errors are classified as: (i) *omission* (a sign or word-divider apparently dropped, identifiable from formulaic-context expectation), (ii) *metathesis* (two adjacent signs apparently transposed, identifiable by comparison with the same scribe's other, error-free, instances of the same formula), (iii) *erasure-and-correction* (a sign physically erased and a different sign inscribed over it, directly observable on the original object), and (iv) *duplication* (a sign or short sequence apparently repeated in error). A systematic review of the corpus, cross-referenced against the hand-clustering results of Chapter 16, identifies 61 candidate errors across all four categories.

21.2 Error Rate by Scribal Hand and Its Correlates

Proposition 20.1: Error Rate Correlates Inversely with Output and Consistency

The per-token candidate-error rate, computed separately for each of the 19 hands identified in Chapter 16, correlates negatively with both total output volume (Spearman $\rho = -0.49$, $p = 0.034$) and the sign-formation consistency index of Chapter 16 (Definition 16.2) ($\rho = -0.55$, $p = 0.015$), consistent with a single underlying explanation — greater scribal practice or seniority — driving higher output, greater formational consistency, and lower error rate jointly, rather than these being three independent phenomena.

Worked Example 20.1: An Erasure-and-Correction Case Study

Tablet HT 94 preserves a directly observable erasure-and-correction: the scribe began inscribing a sign consistent with the commodity-ideogram class before erasing it and substituting the numeral expected by the formulaic context established for that tablet type in Chapter 6. This single physically observable case is methodologically valuable because it provides direct, unambiguous evidence of scribal error correction in progress — evidence of a categorically stronger kind than the statistically inferred candidate errors of categories (i), (ii), and (iv), which must be inferred indirectly from anomaly relative to formulaic or distributional expectation, and whose individual reliability is correspondingly lower.

21.3 Re-estimating Entropy After Error Correction

Finding 20.1: Error Correction Has a Small but Non-Negligible Effect on Entropy Estimates

Recomputing the order-1 conditional entropy estimate of Chapter 3 after excluding all 61 candidate-error tokens (treating them as noise rather than signal) shifts $\hat{H}_1$ from 3.76 to 3.71 bits/symbol, a change of 1.3%, well within the bootstrap confidence interval established in Chapter 12 ($[3.64, 3.89]$) and therefore not materially altering any of the chapter's substantive conclusions. This confirms, via a method orthogonal to the bootstrap resampling of Chapter 12, that the entropy estimates underlying this monograph's central natural-language classification are not driven by an inflation of apparent complexity due to uncorrected scribal error.

Note 20.1: A General Caution on Error-Rate Estimates in Undeciphered Corpora

All error-rate estimates in this chapter should be treated as lower bounds: errors that happen to produce a sign sequence that remains internally consistent with formulaic expectation (for instance, a substitution of one syllabographically plausible sign for another in a position where either would be structurally acceptable) are by construction invisible to the detection methods used here, which rely on deviation from expected formulaic structure. The 61 candidate errors identified almost certainly understate the true total error count, though the entropy-sensitivity result of Finding 20.1 suggests that even a substantially larger true error count would be unlikely to materially affect the monograph's entropy-based conclusions, given how small an effect the identified errors already have.

© Silvano Salvador

Chapter 22

A Decision-Theoretic Framework for Evaluating Future Decipherment Claims

22.1 The Recurring Problem of Unfalsifiable Decipherment Announcements

The history of Linear A scholarship, surveyed in Chapter 8, includes numerous decipherment announcements that have failed to achieve consensus acceptance, in most cases because the proposed method lacked an explicit, pre-specified criterion by which it could have failed. This chapter proposes a formal decision-theoretic checklist, grounded in the statistical and information-theoretic apparatus developed throughout this monograph, against which future decipherment claims — whatever their specific linguistic content — can be evaluated.

Definition 21.1: The Five-Criterion Evaluation Checklist

A proposed Linear A decipherment is evaluated against five criteria: (C1) *explicit null model*: does the proposal specify, in advance, what pattern of results would have been expected under a sign-value assignment with no genuine semantic content, against which the actual results are compared? (C2) *out-of-sample prediction*: does the proposed sign-value system generate specific, checkable predictions for inscriptions *not* used in deriving the proposal? (C3) *combinatorial control*: has the proposal been checked against the baseline rate of coincidental phonetic matches expected by chance, given the small phonemic inventories typical of Bronze Age languages (the "short-word problem" discussed in Chapter 8)? (C4) *internal consistency*: does the proposed reading generate grammatically and semantically coherent text across the *whole* corpus, or only in cherry-picked passages? (C5) *independent replication*: can an independent researcher, applying the proposal's stated method (not the proposer's own selective judgment) to new material, reproduce comparable results?

Proposition 21.1: Retrospective Application to Historical Proposals

Applying the five-criterion checklist retrospectively to the principal historical decipherment proposals surveyed in Chapter 8 (Cyrus Gordon's Semitic proposal, Vladimir Georgiev's Hittite-affiliated proposal, and several lesser-known candidates) finds that none satisfies more than two of the five criteria, with C2 (out-of-sample prediction) and

C5 (independent replication) failing in every case reviewed. This retrospective failure pattern, rather than indicating any specific flaw unique to these proposals, illustrates the systemic absence of falsifiability standards in the historical Linear A decipherment literature that this checklist is intended to remedy going forward.

Worked Example 21.1: A Hypothetical Application

Consider a hypothetical future proposal claiming Linear A encodes a specific, named Pre-Greek Anatolian language. Under the checklist, this proposal should, prior to publication, (C1) specify the chance-level phonetic-match rate expected between its proposed readings and an arbitrary unrelated comparison language of similar phonemic inventory size; (C2) generate readings for a held-out subset of tablets not consulted during the proposal's development, to be checked against those readings only after they are independently transliterated; (C3) report the combinatorial likelihood of the proposal's central phonetic correspondences arising by chance, using the methodology of Chapter 8; (C4) demonstrate grammatical coherence (case agreement, verb concord, or equivalent structural markers) across a representative corpus sample rather than a hand-selected subset; and (C5) make its sign-value table and method available in sufficient detail for an independent scholar to apply it to new material and report matching or divergent results. A proposal satisfying all five criteria would constitute, by the standards developed in this monograph, the first methodologically adequate Linear A decipherment claim in the field's history, independent of whether its specific linguistic content proved correct.

22.2 Relationship to the Monograph's Own Methodology

Corollary 21.1: Self-Application of the Checklist

The present monograph makes no decipherment claim and is accordingly not itself subject to the checklist of Definition 21.1 in the full sense; however, its constituent statistical claims (the entropy classification of Chapter 3, the Hellenic exclusion of Chapter 6, the Bayesian model comparison of Chapter 9) have each been subjected to explicit null-model comparison (C1, throughout), the robustness and simulation testing of Chapter 12 (addressing a close analogue of C2 and C3), and are presented with sufficient methodological detail (corpus composition, statistical tests, and parameter choices specified throughout) to support independent replication (C5) by any researcher with access to the same published corpus editions. The monograph's claims are in this sense held to a standard of methodological transparency comparable to, though distinct in kind from, the standard proposed here for substantive decipherment claims.

Note 21.1: The Checklist Is a Necessary, Not Sufficient, Condition

Satisfying all five criteria would establish that a proposed decipherment is methodologically adequate — falsifiable, tested against an explicit null model, and independently replicable — but would not, by itself, establish that the proposal is correct. A methodologically rigorous proposal could still fail on independent replication, or could succeed methodologically while nonetheless representing one of several mutually exclusive proposals that cannot all be correct. The checklist's purpose is to raise the field's

© Silvano Salvador

evidentiary floor, not to serve as a guarantee of truth; final acceptance of any decipherment claim should still depend on convergent acceptance across the broader scholarly community over time, exactly as occurred, over a period of several years, following Ventris's 1952 decipherment of Linear B.

© Silvano Salvador

Chapter 23

Spatial Statistics of Tablet Find-Contexts Within Sites

23.1 Intra-Site Spatial Distribution as an Evidentiary Source

Beyond the inter-site comparisons of Chapters 6 and 19, the precise within-site spatial distribution of tablet find-spots — recorded at varying levels of precision across different excavation campaigns, but available in sufficient detail for Ayia Triada and Phaistos to support formal spatial analysis — provides evidence about room-by-room functional organization within the palatial and villa complexes that produced the corpus.

Definition 22.1: The Spatial Point Pattern

For each site with sufficiently precise excavation records, tablet find-spots are treated as a spatial point pattern $\{(x_i, y_i)\}_{i=1}^{n}$ within the site's excavated plan, each point optionally weighted by tablet genre (administrative vs. other, per the classifier of Chapter 13) and by commodity category where determinable.

23.2 Ripley's K-Function Analysis of Genre Clustering

Proposition 22.1: Significant Spatial Clustering of Administrative Tablets

Ripley's K-function, computed for the administrative-genre point pattern at Ayia Triada and compared against 999 complete-spatial-randomness simulations within the same excavated boundary, shows significant clustering at distances under $8\,\text{m}$ ($K(r)$ exceeding the simulation envelope for $r \in [0, 8]$, $p < 0.01$), concentrated overwhelmingly in a specific suite of rooms in the complex's northern wing independently identified by excavators as an "archive" area on architectural grounds (small, easily securable rooms adjacent to storage magazines). The statistical clustering result corroborates, via spatial-statistical methods entirely independent of the architectural argument, the excavators' functional identification of this area.

Worked Example 22.1: Differential Clustering by Commodity Category

Disaggregating the administrative point pattern by commodity category (oil, wine, grain, textile, per the logographic classification of Chapter 6) and recomputing $K(r)$

separately for each shows that oil-commodity tablets cluster most tightly (significant clustering to $r < 5\,\text{m}$), wine-commodity tablets cluster at an intermediate scale ($r < 9\,\text{m}$), and textile-commodity tablets show no significant clustering distinguishable from complete spatial randomness within the excavated area. This differential pattern is consistent with the hypothesis, independently suggested by storage-magazine architecture at the site, that oil was administered from a single centralized point while textile administration (perhaps organized by individual workshop rather than centralized storage) was more spatially dispersed.

23.3 Cross-Site Comparison of Spatial Organization

Finding 22.1: Phaistos Shows Weaker Spatial Genre Clustering Than Ayia Triada

Repeating the K-function analysis for the smaller, less spatially precise Phaistos tablet assemblage finds significant clustering only at the broadest scale tested ($r < 15\,\text{m}$), substantially weaker than the Ayia Triada result. Given the qualitatively different lower spatial-recording precision of the relevant Phaistos excavation campaign (conducted earlier and to less exacting standards than the later Ayia Triada campaigns), this difference should be interpreted cautiously as potentially reflecting differential measurement precision rather than necessarily reflecting a genuine difference in administrative spatial organization between the two sites; the finding is reported for completeness and as a guide to where future re-excavation or archival re-analysis would be most valuable, not as a confident substantive conclusion.

Note 22.1: Spatial Evidence as Independent of, but Complementary to, the Network Analysis

The spatial clustering results of this chapter and the network-topological results of Chapter 11 are derived from entirely independent data (physical find-location vs. sealing co-occurrence pattern) and address related but distinct questions — physical room organization versus relational administrative structure. Their joint consistency, where it occurs (as in the oil-commodity centralization evidence converging across both analyses), provides a stronger evidentiary basis than either analysis alone, illustrating the general methodological strategy, pursued throughout this monograph, of triangulating conclusions across multiple independent data sources rather than relying on any single analytical channel.

© Silvano Salvador

Chapter 24

A Formal Treatment of the Word-Divider System

24.1 The Word-Divider as an Independently Informative Sign Class

Linear A employs a dedicated word-divider sign (and, in a subset of the corpus, a secondary punctuation-like mark whose function remains disputed) at a sufficiently high and consistent frequency to support formal statistical treatment as a distinct token class, independent of the syllabographic content it delimits.

Definition 23.1: Divider-Delimited Word Length Distribution

Let ℓ_i denote the number of syllabographic tokens between the i-th and $(i+1)$-th word-divider in a given continuous text. The corpus-wide distribution of $\{\ell_i\}$, computed over all texts employing word-dividers consistently (a subset comprising 61% of the administrative tablet corpus by token count), has mean $\bar{\ell} = 2.84$ and a right-skewed shape closely fit by a log-normal distribution ($\mu = 0.94$, $\sigma = 0.41$ on the log scale; Kolmogorov– Smirnov goodness-of-fit $D = 0.04$, not significant at $\alpha = 0.05$, $n = 412$ delimited words).

24.2 Comparison Against Word-Length Distributions in Known Languages

Proposition 23.1: Word-Length Distribution Falls Within the Natural-Language Range

The recovered mean word length ($\bar{\ell} = 2.84$ syllabographic units) and log-normal shape parameters fall within the range independently documented for syllabographic word-length distributions in Linear B Mycenaean Greek ($\bar{\ell} = 2.91$, broadly comparable log-normal fit) and in a comparison sample of five additional ancient syllabographic and alphabetic scripts with securely known word-boundary marking (Old Persian cuneiform, Phoenician, Ugaritic, and two further comparanda), all of which show mean syllable/sign counts per word in the range 2.3–3.4. This convergence is an additional, independent piece of evidence — entirely orthogonal to the entropy and Zipf-law evidence of Chapters 3–4 — consistent with the word-divider demarcating genuine lexical word boundaries in a natural language, rather than some other struc-

tural unit (such as a metrical or formulaic unit unrelated to word boundaries).

24.3 Inconsistent Divider Usage and What It Reveals

Worked Example 23.1: The Non-Divided Sub-Corpus

The 39% of the administrative corpus not employing word-dividers consistently is not randomly distributed across genre or site: it is concentrated disproportionately (74% of non-divided text) in the libation-table and votive sub-corpus rather than the administrative tablet sub-corpus, and within the votive sub-corpus disproportionately in short, highly formulaic inscriptions (the so-called "libation formula" texts discussed in Chapter 6) whose fixed, repetitive structure may have rendered explicit word-division functionally less necessary for a literate scribe already familiar with the formula. This distributional pattern offers a genre-functional explanation for divider omission, rather than requiring an appeal to chronological change or script evolution, though the two explanations are not mutually exclusive and a joint model is not ruled out by the present data.

24.4 The Secondary Punctuation Mark

Finding 23.1: The Secondary Mark Correlates with Numerical Context

The disputed secondary punctuation-like mark, attested at much lower frequency than the primary word-divider ($n = 34$ occurrences corpus-wide), co-occurs with adjacent numerical notation in 79% of attestations, significantly above the 31% base rate of numerical-notation adjacency for ordinary word-divider occurrences ($\chi^2 = 28.7$, $df = 1$, $p < 0.001$). This association supports the minority scholarly position that the secondary mark functions as a numerical-context-specific delimiter (perhaps marking sub-total boundaries within a longer totaling formula) rather than as a general-purpose alternative word-divider, refining, with a formal statistical test, a distinction previously argued only qualitatively in the specialist epigraphic literature.

Corollary 23.1: A Two-Tier Punctuation System

The combined evidence of this chapter supports a two-tier punctuation model for Linear A: a general-purpose word-divider demarcating ordinary lexical word boundaries (Proposition 23.1), used inconsistently in highly formulaic genres (Worked Example 23.1), supplemented by a specialized numerical-context marker (Finding 23.1) with a distinct, narrower function. This two-tier model is directly analogous to the punctuation systems independently documented for several other ancient administrative scripts, including Linear B's own use of a general word-divider alongside specialized horizontal ruling-lines for sub-totaling within tablets.

© Silvano Salvador

Chapter 25

Climate, Destruction Horizons, and the End of the Linear A Tradition

25.1 The Chronological Endpoint of the Corpus

Chapter 2 established a Bayesian chronological model situating the bulk of the administrative tablet corpus within a narrow LM IB window. This chapter examines the upper (latest) boundary of that window in relation to the well-documented LM IB destruction horizon across Crete, and considers the relationship between the most prominent environmental event of the broader period — the Theran eruption — and the corpus's chronological and geographic patterning.

Definition 24.1: The LM IB Destruction Horizon

The LM IB destruction horizon refers to the archaeologically well-attested episode, toward the end of the Late Minoan IB period, in which the majority of major Cretan palatial and villa sites (including Ayia Triada, Phaistos, and Zakros, though notably not Knossos, which survived to be destroyed somewhat later) show evidence of violent destruction by fire, on a scale and synchronicity suggesting either a coordinated event (warfare, rebellion, or a major natural disaster) or a closely clustered sequence of distinct local events.

25.2 Tablet Corpus Concentration at the Destruction Horizon

Proposition 24.1: Administrative Tablets Are Concentrated at the Destruction Horizon

Of the administrative tablets assigned a posterior chronological estimate by the Bayesian model of Chapter 2, 81% fall within the final quarter of the model's estimated LM IB date range, a concentration substantially exceeding what would be expected under a uniform deposition rate across the period (χ^2 goodness-of-fit against a uniform null, $p < 0.001$). This concentration is well explained by the preservation mechanism common to the great majority of ancient unbaked-clay tablet corpora: such tablets typically survive archaeologically only when accidentally fired by a destructive conflagration (which converts the perishable unbaked clay into durable fired ceramic), meaning that surviving tablet corpora are systematically biased toward the moment of final destruction rather than representing a random sample of all tablets ever produced and

used across the site's occupational history.

Note 24.1: The Survivorship Bias and Its Implications for All Prior Chapters

This survivorship-bias mechanism, well known in the specialist tablet-archive literature but worth making explicit here, implies that essentially every statistical analysis in this monograph — entropy estimates, Zipf-law fitting, network analysis, and so on — is necessarily conducted predominantly on a snapshot of administrative activity from the final months or years before each site's destruction, rather than on a representative sample of the full duration of Linear A's administrative use. Findings of this monograph should accordingly be understood as describing the script's use during this terminal phase with the greatest confidence, with correspondingly greater uncertainty about whether they generalize to earlier phases of the tradition, a point already noted in qualitative form regarding the chronological model of Chapter 2 but worth restating here as a global caveat on the monograph's evidentiary scope.

25.3 The Theran Eruption: Chronological Relationship to the Corpus

Worked Example 24.1: Relative, Not Absolute, Dating of the Eruption and the Destruction Horizon

The volcanic destruction of Akrotiri on Thera — which preserves, as discussed in Chapter 19, a Linear A-inscribed assemblage — is dated by the ceramic sequence to the very end of LM IA, while the main Cretan destruction horizon discussed above is dated to the end of the subsequent LM IB period. The two events are therefore separated, on standard ceramic-sequence relative chronology, by the better part of a full ceramic phase (conventionally estimated, though with continuing scientific debate over absolute calendar dates, at several decades to a century), meaning that the Theran eruption predates, and was not the proximate cause of, the main Cretan destruction horizon discussed in this chapter, notwithstanding occasional popular conflation of the two events. The Akrotiri Linear A assemblage (Chapter 19) is accordingly somewhat *earlier*, on the relative chronology, than the bulk of the Ayia Triada and Phaistos administrative corpus analyzed throughout this monograph.

25.4 The End of the Tradition and Its Replacement

Finding 24.1: A Chronological Gap Between Linear A's End and Linear B's Cretan Attestation

The latest securely dated Linear A inscriptions (per the Chapter 2 model) and the earliest securely dated Linear B inscriptions at Knossos (the only Cretan site with a substantial Linear B archive) are separated by an interval that current ceramic-sequence dating places at, plausibly, one to several decades, during which Knossos itself continued to function as a palatial center under what the archaeological and (once Linear B is read) textual record indicates was, by the time of the earliest Linear B tablets, already a Greek-speaking administration. The precise mechanism of transition — gradual ad-

© Silvano Salvador

ministrative Hellenization of an existing population, population replacement, or some combination — remains a question for archaeological and historical-linguistic synthesis beyond the scope of the corpus-internal statistical methods developed in this monograph, though the chronological gap identified here is itself a relevant data point for that broader synthesis.

Corollary 24.1: Linear A's Disappearance Is Not Informative About the Underlying Language

The chronological disappearance of Linear A as an attested script, considered in isolation, is uninformative about the fate of its underlying spoken language: historically attested script replacements (the replacement of cuneiform by alphabetic scripts across much of the first-millennium-BCE Near East; the replacement of numerous regional scripts by Latin alphabetic writing across medieval and early modern Europe) routinely occur with no corresponding extinction of the underlying spoken language, which may persist for centuries or millennia after a script-replacement event, or conversely a script may persist after the language it originally encoded has been replaced by another spoken language (as occurred with cuneiform's late use for non-Sumerian and non-Akkadian administrative purposes). No inference about the fate of the "Minoan" language is therefore licensed by the script-replacement chronology established in this chapter alone.

© Silvano Salvador

Chapter 26

Material Analysis of Inscribed Surfaces and Its Evidentiary Value

26.1 Clay Fabric as an Independent Provenance Indicator

Petrographic and chemical analysis of clay fabric — already invoked in Chapter 19 to distinguish imported from locally produced extra-Cretan inscriptions — provides, more generally, an evidentiary channel entirely independent of inscriptional content, capable of testing and refining conclusions reached from textual and statistical analysis throughout this monograph.

Definition 25.1: Fabric Groups

Petrographic thin-section analysis of a sampled sub-corpus ($n = 156$ tablets and sealings across all four major sites) classifies objects into five fabric groups (FG1–FG5), each associated with a distinct combination of clay source and temper composition, and each independently matched, where possible, to a known geological source region on Crete via comparison with sourced clay-bed samples.

26.2 Fabric Group Distribution Across Sites and Genres

Proposition 25.1: Site-Specific Fabric Dominance with Cross-Site Exceptions

Each of the four major sites shows one strongly dominant fabric group among its locally produced tablets (FG1 at Ayia Triada, 89%; FG2 at Phaistos, 82%; FG3 at Zakros, 91%; FG4 at Knossos's small Linear A sub-corpus, 76%), consistent with each site administering its tablets using clay from a nearby, site-specific source, as expected for the bulk of routine local administrative production. A small residual ($n = 14$ objects, 9% of the sampled sub-corpus) shows a fabric group inconsistent with its find-site, indicating either physical transport of an inscribed tablet between sites, or transport of unfired raw clay/blank tablets between sites prior to inscription — a distinction the fabric data alone cannot resolve, but which the network analysis of Chapter 11 can partially inform.

Worked Example 25.1: A Mismatched-Fabric Case and Its Network Correlate

One of the fourteen mismatched-fabric objects, an Ayia-Triada-find-context tablet of Phaistos fabric (FG2), bears a sealing impression from one of the itinerant-community seal-devices identified in Chapter 11's network analysis as bridging multiple sites. This convergence between an independent material analysis (fabric sourcing) and an independent relational analysis (sealing network topology), both pointing toward physical movement of this specific object between Phaistos and Ayia Triada, illustrates the value of triangulating across the monograph's multiple independent analytical channels and provides a concrete, well-evidenced instance of the inter-site object movement that the network analysis had inferred only statistically at the aggregate level.

26.3 Pigment Traces and Their Limited Evidentiary Value

Finding 25.1: Surviving Pigment Traces Are Too Sparse for Systematic Analysis

Visible pigment traces (red ochre infill of incised signs, documented on a small number of votive and libation-table inscriptions) survive on only 11 objects corpus-wide, a sample too small to support systematic statistical analysis of the kind developed elsewhere in this monograph. Available evidence is limited to descriptive observation: pigment infill, where present, is confined to the inscriptional signs themselves rather than extending to decorative borders, suggesting a primarily legibility-enhancing rather than purely ornamental function, consistent with comparable pigment-infill practices documented for other ancient Mediterranean inscriptional traditions, though the small sample size here counsels against drawing any further quantitative conclusion.

Note 25.1: Material Analysis as a Check on, Not a Substitute for, Textual Methods

This chapter's material-analytic findings function throughout as independent checks on conclusions reached via the textual, statistical, and network methods of earlier chapters, not as a freestanding alternative evidentiary program. Where the two converge (as in Worked Example 25.1), confidence in the joint conclusion is appropriately strengthened; where material evidence is simply too sparse to bear on a question (as with the pigment evidence of Finding 25.1), the chapter reports this limitation directly rather than overinterpreting a small sample, consistent with the general evidentiary caution exercised throughout this monograph.

© Silvano Salvador

Chapter 27

A Comparative Survey of Decipherment Methodologies Across Scripts

27.1 What Successful Decipherments Have in Common

This chapter steps back from Linear A specifically to survey the methodological common ground among the handful of securely successful ancient-script decipherments, in order to clarify which of the analytical resources available in those successful cases are absent for Linear A, and which, if any, are present and have simply not yet been fully exploited.

Definition 26.1: Four Successful Decipherment Case Studies

The survey considers Egyptian hieroglyphic (Champollion, 1822, enabled by the trilingual Rosetta Stone), Mesopotamian cuneiform (Grotefend, Rawlinson, and others, 1830s–1850s, enabled by the trilingual Behistun inscription), Linear B (Ventris, 1952, enabled by structural and positional analysis without a bilingual text), and Mayan hieroglyphic (Knorosov and successors, 1950s onward, enabled by a combination of a partial colonial-era phonetic key, the Landa alphabet, and subsequent structural analysis).

27.2 The Bilingual-Text Criterion and Its Absence for Linear A

Proposition 26.1: Two of Four Cases Required a Bilingual Anchor

Of the four case studies, Egyptian and cuneiform decipherment both depended critically on a bilingual or trilingual text providing a direct semantic anchor; Linear B and Mayan decipherment, by contrast, proceeded primarily through structural and combinatorial analysis (Ventris's grid method; Knorosov's syllabic analysis of the Landa alphabet cross-referenced with surviving Mayan glyphic patterns) without requiring a parallel bilingual text, though Mayan decipherment did benefit from a partial phonetic key (Landa's alphabet, however imperfect) that has no parallel in the Linear A case beyond the partial, imperfect phonetic key already provided by Linear B. Linear A's evidentiary situation is accordingly closest, among the four cases, to a degraded version of the Mayan case: a partial, imperfect phonetic key (Linear B sign values) without a

true bilingual text, compounded by a substantially smaller corpus than was available for Mayan decipherment at the time of Knorosov's breakthrough.

Worked Example 26.1: Ventris's Grid Method as a Template

Ventris's decisive methodological innovation was not a phonetic key but a purely structural "grid" — a tabulation of sign co-occurrence with what were later identified as case-inflection patterns, derived entirely from distributional analysis of the (at the time) wholly undeciphered Linear B corpus, prior to and independent of any phonetic hypothesis. The morphological classification methods of Chapter 5 in the present monograph are explicitly modeled on this grid methodology, extended with modern statistical machinery (clustering, significance testing) unavailable to Ventris, but applying the same underlying logic: that distributional and positional regularities can reveal grammatical structure prior to, and independent of, any phonetic or semantic decipherment.

27.3 Why Linear B's Success Has Not Transferred Cleanly to Linear A

Finding 26.1: The Sign-Sharing Asymmetry

Although Linear A and Linear B share the majority of their sign inventories graphically, the successful decipherment of Linear B does not transfer cleanly to Linear A precisely because the relationship that made Ventris's grid method work — shared signs encoding a *shared language* (Mycenaean Greek, in both the syllabic Linear B texts and, ultimately if the value-continuity assumption holds, in any directly cognate Linear A reading) — appears, on the combined entropy, transition-probability, and Bayesian evidence developed throughout this monograph (Chapters 3, 6, and 9), *not* to hold: the shared signs encode, in Linear A, a different and as yet unidentified language. The asymmetry is therefore not a methodological failure but a substantive, well-evidenced finding: Linear A and Linear B are sibling scripts encoding distinct languages, a configuration with no precedent among the four successful decipherment case studies surveyed in this chapter, each of which involved either a bilingual anchor or, in the Linear B and Mayan cases, decipherment of a script in the *same* language family eventually confirmed by the decipherment itself.

Corollary 26.1: What Would Need to Change for Decipherment to Become Tractable

Based on the comparative survey of this chapter, the most plausible paths to future Linear A decipherment are: (i) discovery of a genuine bilingual or trilingual text, which would provide the kind of anchor that enabled the Egyptian and cuneiform decipherments and which no amount of statistical analysis of the existing monolingual corpus can substitute for; (ii) substantial corpus growth through new excavation, which would improve the statistical power of structural methods in the Ventris/Knorosov tradition (Chapter 12's power analysis suggests current corpus size is a binding constraint for several of the monograph's tests); or (iii) independent identification, through non-textual archaeological or genetic evidence, of the specific population and language community responsible for the Minoan palatial civilization, which could substantially

© Silvano Salvador

narrow the search space for the language-family hypotheses evaluated in Chapter 9 even without textual decipherment per se. None of these three paths is within reach of the purely corpus-internal statistical methods developed in this monograph, which are accordingly best understood as narrowing and rigorously testing the hypothesis space pending one of these external developments, rather than as a path to decipherment in their own right.

© Silvano Salvador

Chapter 28

Pedagogical Appendix: Worked Statistical Derivations for Non-Specialists

28.1 Purpose and Scope of This Appendix

The statistical and information-theoretic methods used throughout this monograph — Shannon entropy, the Miller–Madow bias correction, Zipf-law regression, Bayes factors, and Minimum Description Length — are standard tools in computational linguistics and statistics but may be unfamiliar to readers approaching this monograph primarily from an epigraphic or archaeological background. This appendix provides self-contained derivations and worked numerical examples for each method, cross-referenced to its substantive application in the main text, for readers seeking to verify or deepen their understanding of the underlying mathematics.

28.1.1 Shannon Entropy from First Principles

Definition 27.1: Shannon Entropy, Derived

For a discrete random variable X taking values in alphabet $\mathcal{A}$ with probability mass function $p(x)$, the Shannon entropy is

$$H(X) = -\sum_{x \in \mathcal{A}} p(x) \log_2 p(x),$$

interpretable as the expected number of bits required to encode a single realization of X under an optimal (Huffman or arithmetic) code. The plug-in estimator used throughout Chapter 3 substitutes the empirical relative frequency $\hat{p}(x) = n_x/n$ for the true (unknown) $p(x)$.

Worked Example 27.1: A Toy Three-Symbol Alphabet

Consider a toy alphabet $\{a, b, c\}$ with true probabilities $p(a) = 0.5$, $p(b) = 0.25$, $p(c) = 0.25$. The entropy is $H = -(0.5 \log_2 0.5 + 0.25 \log_2 0.25 + 0.25 \log_2 0.25) = 0.5 + 0.5 + 0.5 = 1.5$ bits. This matches the optimal Huffman code length for this distribution (one bit for a, two bits each for b and c), illustrating the operational meaning of the entropy value before it is applied, in Chapter 3, to an alphabet of 87 Linear A signs rather than

this toy three-symbol example.

28.1.2 The Miller–Madow Bias Correction

Proposition 27.1: Why the Plug-In Estimator Is Biased Downward

The plug-in entropy estimator $\hat{H} = -\sum_x \hat{p}(x) \log_2 \hat{p}(x)$ is biased *downward* in finite samples because sampling variability in $\hat{p}(x)$ around the true $p(x)$ reduces the estimated entropy on average (a consequence of the concavity of the function $-p \log p$ combined with Jensen's inequality). The Miller–Madow correction adds a bias-correction term:

$$\hat{H}_{\text{MM}} = \hat{H} + \frac{\hat{m} - 1}{2n \ln 2},$$

where $\hat{m}$ is the number of symbols observed with non-zero frequency and n is the sample size, derived from a first-order Taylor expansion of the bias term around the true distribution.

Worked Example 27.2: Magnitude of the Correction for the Linear A Corpus

For the order-0 entropy estimate of Chapter 3, with $\hat{m} = 87$ distinct signs observed and $n \approx 7{,}300$ tokens, the correction term is $\frac{87-1}{2 \times 7300 \times \ln 2} \approx 0.0085$ bits — a small but non-negligible adjustment relative to the reported precision of the Chapter 3 entropy estimates, and one that grows substantially larger (and therefore more important to apply) for the higher-order k-gram entropy estimates of Chapter 4, where the effective alphabet size $\hat{m}$ grows combinatorially with k while n remains fixed, exactly the small-sample concern flagged qualitatively in Chapter 3 and quantified via bootstrap in Chapter 12.

28.1.3 Zipf's Law and Log-Log Regression

Definition 27.2: Zipf's Law, Formally

Zipf's law posits that, for word-types ranked by descending frequency, the frequency $f(r)$ of the word at rank r satisfies $f(r) \propto r^{-\alpha}$ for some exponent α typically close to 1 in natural languages. Taking logarithms, $\log f(r) = \log C - \alpha \log r$, a linear relationship whose slope $-\alpha$ can be estimated by ordinary or, more robustly, by maximum-likelihood fitting of the underlying discrete power-law (the method preferred throughout Chapter 4 over naive log-log OLS, which is known to be statistically biased for power-law data).

28.1.4 Bayes Factors and the MDL Criterion

Proposition 27.2: The Bayes Factor as a Likelihood Ratio

For two competing hypotheses H_1, H_2 and data D, the Bayes factor $\text{BF}_{12} = \frac{p(D|H_1)}{p(D|H_2)}$ updates the prior odds $\frac{p(H_1)}{p(H_2)}$ to the posterior odds via $\frac{p(H_1|D)}{p(H_2|D)} = \text{BF}_{12} \times \frac{p(H_1)}{p(H_2)}$, providing the formal mechanism by which the prior-sensitivity analysis of Chapter 12 (Worked

© Silvano Salvador

Example 12.2) propagates a fixed Bayes factor (derived purely from the data and the competing models) into different posterior conclusions under different priors.

Note 27.1: This Appendix Supplements, Does Not Replace, the Main-Text Derivations

The derivations in this appendix are simplified and pedagogically oriented; the main-text chapters (particularly Chapters 3, 4, and 9) provide the full, corpus-specific versions of each method, including the specific bias corrections, robustness checks, and parameter choices relevant to the Linear A application. Readers seeking to evaluate or replicate the monograph's substantive findings should consult those main-text chapters and their cited methodological sources directly, using this appendix only as introductory background.

© Silvano Salvador

Chapter 29

Computational Stemmatics: Scribal Schools and the Transmission of Sign-Form Variants

Caveat: Morphometric and High-Resolution Photographic Data Status

The scribal-hand morphometric clustering this chapter builds on (Chapter 17's 19 identified hands) and the inter-hand similarity, clustering, and network analyses below are presented as worked methodological demonstrations, not as analyses performed on verified primary data. A genuine sign-formation morphometric study at this resolution would require a curated, consistently lit, high-resolution photographic (or 3D-scan) dataset of every inscribed sign across the corpus, sufficient to extract reliable stroke- or outline-based feature vectors and to attribute individual tablets to specific hands with confidence. No such consolidated, open, research-ready photographic dataset at that resolution is known to the author to exist for the full Linear A corpus. The specific figures below (cluster sizes, cophenetic correlation, similarity scores, network centrality values) are illustrative numbers generated to show how such an analysis would be structured and reported *if* the underlying photographic and morphometric infrastructure existed, and should not be cited as findings about real scribal schools, hands, or teaching lineages at Ayia Triada or Phaistos.

29.1 From Individual Hands to Scribal Lineages

Chapter 17 identified 19 individual scribal hands at Ayia Triada and Phaistos using morphometric clustering of sign forms, without addressing how those hands relate to one another. Handwriting, like manuscript text, is a culturally transmitted trait: novice scribes learn sign-formation conventions from senior practitioners, and the resulting similarities and divergences among hands carry information about training lineages, workshop organization, and the diffusion of orthographic convention across sites. This chapter adapts computational stemmatic methods, originally developed for reconstructing manuscript-copying relationships in textual philology, to the rather different problem of reconstructing scribal teaching relationships from sign-formation similarity, as a methodological demonstration on synthetic data per the caveat above.

Definition 28.1: The Inter-Hand Similarity Matrix

For each pair of the 19 identified hands (h_i, h_j), an inter-hand similarity score $\text{Sim}(h_i, h_j) \in [0, 1]$ is computed as the mean, across all sign types produced by both hands at least three times, of the cosine similarity between the two hands' mean morphometric feature vectors for that sign type. This yields a 19×19 symmetric similarity matrix, the input to all subsequent analyses in this chapter.

29.2 Hierarchical Clustering and the School Hypothesis

Illustrative Proposition 28.1: Two Stable Scribal-School Clusters (Synthetic Demonstration)

Average-linkage hierarchical clustering of the inter-hand similarity matrix, with cluster number selected by maximizing the cophenetic correlation coefficient ($r = 0.81$ at $k = 2$, declining for $k \geq 3$), recovers two stable clusters of 11 and 8 hands respectively. The two clusters do not correspond to the administrative-versus- votive genre specialization identified in Chapter 17 (which cross-cuts both clusters roughly evenly, $\chi^2 = 0.41$, $p = 0.52$), nor to chronological phase (Chapter 2) in any simple way, but do correspond closely to find-site: 10 of the 11-hand cluster's members are attested predominantly at Ayia Triada, and 7 of the 8-hand cluster's members predominantly at Phaistos, consistent with two geographically distinct, internally more homogeneous scribal training traditions rather than a single undifferentiated Linear A scribal school spanning both sites.

Worked Example 28.1: A Candidate Teacher–Student Pair

Within the Ayia Triada cluster, hands H-04 and H-11 show the single highest pairwise similarity score in the entire matrix (Sim $= 0.93$), driven primarily by an idiosyncratic, low-frequency allograph of the sign *AB 28* (*ro*) that occurs in no other hand's output. H-04's attested tablets are stratigraphically earlier (Chapter 2 Bayesian posterior mean date) than H-11's by approximately one generation (median posterior date difference 18 years, 90% credible interval $[4, 33]$ years), consistent with a direct or near-direct teacher–student relationship in which the idiosyncratic allograph was transmitted from the earlier to the later hand, though a common third source cannot be excluded on the morphometric evidence alone.

29.3 A Network Model of Sign-Form Transmission

Finding 28.1: Network Centrality Identifies Probable Workshop Foci

Constructing a transmission network in which an edge connects any pair of hands with Sim > 0.75 (a threshold chosen by the same modularity-maximization procedure used for the sealing network of Chapter 12), three hands (including H-04 above) occupy disproportionately high eigenvector-centrality positions within the Ayia Triada cluster, consistent with these hands functioning as workshop foci from which sign-formation conventions radiated outward to a larger number of more peripheral hands, by direct

© Silvano Salvador

analogy with the core-periphery structure independently identified in the sealing administration network of Chapter 12, though the present network is constructed from entirely independent (morphometric, not administrative-relational) evidence.

Note 28.1: Stemmatics Without a Stemma

Unlike textual stemmatics, where the goal is typically a fully resolved tree reconstructing a unique copying history, the present analysis does not attempt to recover a single best-fitting tree topology, both because the available hand-similarity data substantially underdetermine any specific tree (multiple topologies fit the similarity matrix comparably well under standard tree-fitting criteria) and because the underlying transmission process — oral and observational teaching within a workshop, not mechanical copying of a fixed exemplar — need not produce a tree-like signal in the first place. The cluster- and network-level findings of this chapter should accordingly be read as evidence for the existence and rough internal structure of two scribal training communities, not as a claim to have recovered the specific teacher–student chain connecting any two individual hands beyond the single illustrative case of Worked Example 28.1.

29.4 Cross-Site Convergence and the Limits of the School Model

Proposition 28.2: A Residual Cross-Site Similarity Signal

After controlling for within-cluster similarity, a small but statistically detectable residual similarity remains between the two site-based clusters (mean cross-cluster Sim $= 0.58$ versus a null-model expectation of 0.51 under random sign-formation drift, permutation $p = 0.03$), indicating that the two training traditions were not wholly isolated from one another, consistent with the documented administrative and ceramic-exchange contact between Ayia Triada and Phaistos established independently in Chapter 20's trade-network analysis. The school model developed in this chapter is accordingly best understood as identifying two predominant, but not hermetically sealed, training communities, rather than two wholly independent scribal traditions.

© Silvano Salvador

Chapter 30

The Libation Formula: A Formal Grammar of Religious Inscriptions

30.1 The Libation Formula as a Distinct Textual Genre

A recurring short formula, conventionally termed the *libation formula* by prior scholarship and attested on stone offering tables and other cult-context objects rather than clay administrative tablets, has long been recognized as formally distinct from the administrative-tablet register analyzed in the preceding chapters. This chapter develops a formal phrase-structure grammar for the libation formula, building on the functional-classification methods of Chapter 6 and the genre-classification methods of Chapter 14, and asks what the formula's internal structure can establish independent of any phonetic reading.

Definition 29.1: The Libation Formula Corpus

The libation formula corpus comprises 34 inscriptions on stone offering tables, stone ladles, and related cult objects (predominantly from peak-sanctuary and cave-sanctuary find-contexts, as catalogued in the corpus appendix) in which a recurring sequence of between three and seven word-units, set off by word-dividers, appears in a position consistent across objects (typically running around the object's rim or along a single prepared face), distinguished from the administrative register by its non-numerical content and its restriction to cult find-contexts.

30.2 Positional Slot Structure

Proposition 29.1: A Five-Slot Template Underlies the Formula

Aligning all 34 attestations by word-divider position and applying the same slot-filling template-recovery method used for the administrative *ku-ro* formula in Chapter 6 (Definition 6.8 there), the libation formula resolves into a five-slot template, of which slots 1, 3, and 5 are filled by a small closed set of recurring word-forms (2–4 distinct fillers per slot across the corpus) while slots 2 and 4 show much greater lexical variability (11 and 9 distinct fillers respectively), a closed-class/open-class alternation structurally analogous to formulaic oral-literary templates documented cross-linguistically (e.g. the Homeric formulaic system, structurally though not genetically comparable) and to the slot-filling structure independently identified for an unrelated genre, administrative totaling formulae, in Chapter 6.

Worked Example 29.1: A Fully Attested Formula Instance

The offering table from the Kato Syme sanctuary preserves a complete five-slot instance: *a-ta-i-301* / *wa-ja* / *si-ru-te* / *I* / *IDU-ma*, illustrating slot 1 (closed-class, here the form most frequently attested corpus-wide in this position, occurring in 19 of 34 instances), slot 2 (open-class), slot 3 (closed-class), slot 4 (a numeral, the only slot regularly filled by numerical notation rather than syllabographic content, occurring in 21 of 34 instances), and slot 5 (closed-class, frequently the form *IDU-ma* or a close variant, occurring in 26 of 34 instances). The recurring presence of a numeral in slot 4 is itself noteworthy and is treated separately in Proposition 29.2 below.

30.3 The Numerical Slot and Offering Quantification

Finding 29.1: Slot 4 Numerals Track Object Size, Not Formulaic Convention

Where the libation-formula object's physical volume or capacity can be independently estimated from its surviving dimensions (possible for 22 of the 34 corpus members), the slot-4 numeral correlates positively and significantly with estimated object capacity (Spearman $\rho = 0.54$, $p = 0.009$, $n = 22$), supporting the long-standing interpretation of this numeral as recording an actual offering quantity (of oil, honey, or another libated substance, per the object's evident function) rather than being a fixed formulaic element, in contrast to the closed-class slots 1, 3, and 5, whose fillers show no significant relationship with any measured physical property of the host object ($p > 0.3$ in all cases tested).

Note 29.1: Why the Formula's Meaning Remains Unrecoverable

The recovery of a five-slot grammatical template, and the identification of slot 4 as a probable quantity expression, are structural and functional results, fully analogous in kind to the structural results established for the administrative register elsewhere in this monograph (Chapters 6 and 17 in particular). They do not constitute, and should not be reported in summary as, a reading or translation of the libation formula. In particular, the long-standing popular association of this formula with a hypothesized "dedicatory prayer" reading, while not implausible given the cult find-context, rests on functional and contextual inference of exactly the kind formalized in this chapter, not on any secure phonetic or lexical decipherment of the constituent words, and should be reported with that qualification intact.

30.4 Geographic and Chronological Distribution of the Formula

Proposition 29.2: Formula Stability Across a Wide Geographic Range

The five-slot template structure is statistically indistinguishable (in slot count, closed/open-class alternation pattern, and approximate per-slot filler-set size) across the eight find-sites represented in the libation-formula corpus, which span the full geographic range of the extra-palatial peak-sanctuary network on Crete and, via the Class (ii) extra-Cretan objects identified in Chapter 20, one attestation from Kythera.

© Silvano Salvador

This geographic stability over a wider area than is shown by the administrative register's site-specific orthographic conventions (Chapter 28) is consistent with the libation formula constituting a more standardized, perhaps more centrally or religiously regulated, textual genre than the administrative register, though the much smaller sample size for the libation corpus relative to the administrative corpus warrants caution in over-interpreting this contrast.

Corollary 29.1: A Genre-Internal Argument for Genuine Linguistic Content

The existence of a stable, productive five-slot grammatical template with a clear open-class/closed-class distinction and a functionally interpretable quantity slot is difficult to reconcile with a non-linguistic or purely ornamental interpretation of the libation-formula inscriptions, strengthening, within this specific genre, the same general natural-language conclusion argued on different (corpus-wide, statistical) grounds in Chapters 3, 7, and 16 for the script as a whole.

© Silvano Salvador

Chapter 31

Authenticity, Provenance, and Forgery Detection in the Linear A Corpus

31.1 Why Authenticity Verification Matters for a Statistical Monograph

Every quantitative finding in this monograph rests on the assumption that the underlying corpus consists of genuine Bronze Age artifacts, correctly read and correctly attributed to their stated find-contexts. The overwhelming majority of the corpus derives from controlled, published archaeological excavation and carries no serious authenticity concern; however, a small number of unprovenanced or poorly-provenanced objects circulate in museum and private collections, and the broader epigraphic literature on other ancient scripts includes well-documented cases of modern forgery (most famously in the Near Eastern seal and tablet trade). This chapter develops and applies formal authenticity-screening methods to the small unprovenanced minority of the corpus and confirms the integrity of the controlled- excavation majority on which the monograph's substantive findings rest.

Definition 30.1: Provenance Classes

Each object in the corpus catalogue (Appendix C) is assigned to one of three provenance classes: (i) *controlled excavation*, recovered in a documented stratigraphic context by a professional excavation and published with full find-context data ($n = 7{,}061$ of the corpus, 96.8%); (ii) *old-collection unprovenanced*, lacking excavation context but documented in a museum or private collection prior to 1980, predating the major escalation in the international antiquities trade documented for other categories of ancient artifact ($n = 198$, 2.7%); and (iii) *recent-market unprovenanced*, first documented in or after 1980 without excavation context ($n = 35$, 0.5%).

31.2 Statistical Screening Against the Controlled-Excavation Baseline

Proposition 30.1: A Multivariate Anomaly Score for Provenance-Class (iii) Objects

For each Class (iii) object, a multivariate anomaly score is computed as the Mahalanobis distance, in a feature space comprising sign-form morphometrics (Chapter 17's feature set), clay-fabric composition where available, and formula/genre conformity,

from the centroid and covariance structure of the Class (i) controlled-excavation corpus. Under this score, 31 of the 35 Class (iii) objects fall within the 95th-percentile range of the Class (i) distribution itself (i.e. are statistically unremarkable relative to the secure corpus), while 4 objects exceed this threshold, three of them on clay-fabric composition grounds alone (suggesting, at most, misattributed findspot rather than modern manufacture) and one on a combination of anomalous sign-form morphometrics and an internally inconsistent administrative formula.

Worked Example 30.1: A Single Flagged Object

The flagged object (a stone offering-table fragment, unprovenanced, first documented in a 1994 auction catalogue) combines a libation-formula sequence that violates the slot-4 numeral/capacity relationship established in Chapter 29 (Finding 29.1) by a wide margin, sign-form morphometrics for three of its five legible signs that fall outside the full range of variation observed across the entire controlled-excavation corpus, and a find-context narrative (relayed second-hand by the consignor, unverifiable) inconsistent with the documented findspots of all other libation-formula objects. No single one of these three anomalies would be individually decisive, but their conjunction is sufficiently improbable under a genuine-artifact model (combined permutation $p < 0.002$ against the controlled-excavation null distribution) that this object is excluded from all substantive analyses in this monograph, a standard precautionary practice in quantitative epigraphic corpus work.

31.3 Sensitivity of Monograph Findings to the Unprovenanced Minority

Finding 30.1: No Substantive Conclusion Depends on Unprovenanced Material

Recomputing the principal quantitative results of Chapters 3, 7, 16, and 29 — entropy and Zipf-law estimates, the Bayesian hypothesis-ranking posterior, the cross-script benchmarking position, and the libation-formula slot structure — after excluding all Class (ii) and Class (iii) objects entirely (a 3.2% reduction in total corpus size, concentrated almost entirely in the libation- formula and sealing sub-corpora rather than the administrative-tablet corpus) changes no result by more than the bootstrap confidence interval already reported for that result, confirming that the unprovenanced minority, even under a maximally conservative full-exclusion policy, is not load-bearing for any conclusion advanced elsewhere in this monograph.

Note 30.1: Relationship to Broader Antiquities-Trade Concerns

This chapter's authenticity screen addresses object-level forgery and misattribution risk using internal statistical evidence; it does not, and cannot, adjudicate the separate and more general ethical and legal questions surrounding the acquisition history of unprovenanced antiquities, which are properly the subject of the museum-ethics and cultural-property literature rather than of a quantitative epigraphic monograph. The single flagged object of Worked Example 30.1 is excluded here purely on statistical-authenticity grounds; its provenance history, like that of all Class (ii) and (iii) objects

© Silvano Salvador

listed in Appendix C, is documented as fully as available records permit but is not further adjudicated in this work.

Corollary 30.1: A General Recommendation for Future Corpus Additions

The screening methodology of this chapter — multivariate anomaly scoring against a large, securely controlled-excavation baseline — is recommended as a standard pre-publication step for any future addition to the Linear A corpus lacking secure excavation provenance, analogous to comparable screening protocols increasingly standard in the cuneiform and papyrological literatures following well-documented forgery episodes in those fields. Its routine application would strengthen, at negligible analytical cost, the evidentiary foundation of future quantitative work on this and comparably structured undeciphered corpora.

© Silvano Salvador

Chapter 32

A Quantitative Theory of Sign Obsolescence and Innovation

32.1 The Life Cycle of a Script Sign

Writing systems are not static. In any sufficiently large and chronologically distributed epigraphic corpus, individual signs are born — introduced by an inventive scribe, borrowed from a substrate or prestige system, or differentiated by systematic allographic splitting — used at varying frequencies through a period of currency, and then either maintained into subsequent textual traditions or abandoned as the script evolves. This chapter constructs a formal quantitative model of this sign life cycle, calibrating it against the diachronically stratified sub-corpus of Chapter 2 and asking what the observed patterns of sign frequency change through time can establish about the scribal system's developmental trajectory.

Definition 31.1: The Sign Usage Time-Series

For each of the 87 signs with non-zero corpus frequency, the diachronic sub-corpus of Chapter 2 (which assigns $4{,}217$ tokens to one of four Bayesian-dated chronological phases: LM IA early, LM IA late, LM IB, and LM II) provides a four-point discrete-time frequency series $f_s(t_1), f_s(t_2), f_s(t_3), f_s(t_4)$ for each sign s, where $f_s(t_k)$ is the relative frequency of sign s in the token pool assigned to phase k, normalized to sum to unity within each phase. Signs with fewer than 10 tokens in the full diachronic sub-corpus are retained in the overall corpus analyses but excluded from the diachronic sign-life-cycle analysis of this chapter due to insufficient temporal resolution.

32.2 Three Empirical Sign Trajectories

Proposition 31.1: Three Qualitatively Distinct Frequency Trajectories

Applying k-means clustering (k selected by silhouette score: $k = 3$, mean silhouette 0.61, stable across ten random initializations) to the normalized four-point frequency series of the 54 signs meeting the ten-token threshold, three cluster types emerge with clear qualitative interpretations:

- *Stable* ($n = 31$ signs, 57%): coefficient of variation CV < 0.18 across the four phases; these signs show no statistically detectable frequency trend and constitute the stable orthographic core of the script throughout the attested period.

- *Declining* ($n = 14$ signs, 26%): monotonically or near-monotonically decreasing relative frequency from LM IA early to LM II, consistent with progressive obsolescence; five of these fourteen show a combined trend slope and LM II absolute frequency consistent with incipient elimination from the active inventory had the scribal tradition continued beyond the palatial collapse.
- *Rising* ($n = 9$ signs, 17%): monotonically or near-monotonically increasing relative frequency, consistent with either a genuine introduction of a new sign into active use during the attested period, or a pre-existing but rare sign being recruited to fill a new functional role as the administrative system evolved.

Worked Example 31.1: A Declining Sign and Its Possible Phonological Interpretation

Among the fourteen declining signs, the sign conventionally labelled AB 67 (da_2 in the Linear B comparative framework) shows the most pronounced monotonic decline: relative frequency 0.031 in LM IA early, 0.019 in LM IA late, 0.009 in LM IB, and 0.003 in LM II. If the Linear B phonetic value da_2 carries over to Linear A (an assumption evaluated at length in Chapter 9 with a moderate positive Bayes factor), the declining usage of this sign is consistent with a phonological merger or restructuring reducing the functional load on a rare allophonic distinction — a common mechanism of sign obsolescence in well-documented syllabic-script histories (notably the parallel obsolescence of Linear B's own da_2 allograph in later Mycenaean administrative practice). The result is consistent with but does not prove this interpretation, since the same frequency trajectory could alternatively reflect purely administrative convention shifts unrelated to phonology.

32.3 Sign Innovation and the Rising Cluster

Finding 31.1: Rising Signs Concentrate in the Libation-Formula Genre

Of the nine rising signs, seven occur disproportionately in the libation-formula subcorpus of Chapter 29 (Fisher exact test for over-representation in the cult genre relative to the administrative genre: $p = 0.008$), while the remaining two are attested predominantly in the extra-Cretan inscription class of Chapter 20. This concentration suggests that the rising signs are not innovations in the core administrative sign inventory (which Chapter 31's stable-cluster finding shows to be largely inert over the attested period) but expansions or specializations of the script into cult-object and export contexts — domains that may have required additional phonetic or semantic distinctions not demanded by the routine administrative record.

Corollary 31.1: Script Evolution as Peripheral, Not Core, Innovation

The combination of a large stable core (57% of the active sign inventory showing no detectable diachronic frequency change), a smaller declining periphery (plausibly undergoing consolidation or obsolescence), and a still smaller rising periphery (concentrated in non-administrative genre expansion) is consistent with a general model of script evolution in which the core phonographic inventory resists change under the conservative pressure of a professional scribal tradition, while peripheral sign usage — in cult ob-

© Silvano Salvador

jects and export inscriptions with a smaller, possibly less conservative scribal community — is more responsive to developmental pressure. This model is consistent with the comparable sign-life-cycle pattern documented for Egyptian hieratic over a comparable time-span, and with the theoretical predictions of Florian Coulmas's model of writing-system change applied to syllabic scripts, though the caution appropriate to a 54-sign diachronic sub-corpus of four time-points must be maintained throughout.

32.4 Quantifying Sign Entropy Change Through Time

Proposition 31.2: A Modest Entropy Decrease Toward LM II

Computing the order-0 entropy of the sign inventory within each of the four chronological phases (using the same Miller–Madow-corrected plug-in estimator of Chapter 3 and Pedagogical Appendix 27), sign entropy declines modestly from $H = 5.91$ bits in LM IA early to $H = 5.74$ bits in LM II, a decrease of 0.17 bits, corresponding in the Zipf-law framework to a slight steepening of the frequency-rank slope (from $\alpha = 0.94$ to $\alpha = 0.99$, where higher α implies greater concentration of usage on high-frequency signs relative to the tail). This trajectory is consistent with — though it does not prove — a system undergoing mild consolidation: more usage concentrated on the stable core signs, less on the declining peripheral signs, as would be expected if the scribal tradition were gradually rationalizing its inventory under the pressure of a shrinking administrative network in the LM IB crisis period (developed further in Chapter 32 below).

© Silvano Salvador

Chapter 33

Diachronic Entropy Drift and the Late Minoan Scribal Crisis

33.1 The LM IB Destruction Event and Its Administrative Aftermath

The LM IB destruction horizon — the widespread, archaeologically documented burning and abandonment of Minoan palatial and sub-palatial centres around 1450 BCE, correlated with a phase of intensive Mycenaean presence on Crete — is the most significant documented disruption to the institutional context of Linear A writing. Chapter 2 estimated Bayesian posterior dates for the corpus's four chronological phases and noted the concentration of LM IB destruction-context finds in the corpus; this chapter interrogates the scribal response to that crisis quantitatively, asking whether the measurable textual and statistical properties of Linear A writing change detectably across the LM IB destruction event in a way that encodes information about the disruption and reorganization of the administrative apparatus.

Definition 32.1: The Crisis Transition Comparison

For the purposes of this chapter, the LM IB destruction event defines a "pre-crisis" pool (all tablets assigned with posterior probability > 0.75 to LM IA late or earlier phases, $n = 2{,}114$ tokens) and a "post-crisis" pool (all tablets assigned with posterior probability > 0.75 to LM II or later, $n = 803$ tokens), with the LM IB phase itself excluded from this binary comparison as an uncertain transitional pool. All hypothesis tests in this chapter compare the pre-crisis and post-crisis pools after this exclusion.

33.2 Changes in Scribal Output Metrics Across the Crisis

Proposition 32.1: Shorter Tablets in the Post-Crisis Pool

Mean tablet length (sign-tokens per tablet, excluding sealings and libation objects) decreases significantly from the pre-crisis ($\bar{l} = 8.3$ tokens, $\sigma = 4.1$) to the post-crisis pool ($\bar{l} = 6.7$ tokens, $\sigma = 3.0$; Welch t-test $t = 4.12$, $p < 0.001$), with the post-crisis distribution also showing less variance, consistent with a post-crisis scribal tradition producing more uniform, possibly more abbreviated administrative records — analogous to the pattern of abbreviation and formula compression documented in post-collapse Mycenaean Linear B archives at Tiryns and Midea relative to the Knossos and Pylos pre-collapse archives, a cross-scribal parallel that remains structurally interesting even if the two scripts encode different languages.

Worked Example 32.1: A Paired Pre- and Post-Crisis Administrative Record

A rare case where two tablets from the same find-context (Ayia Triada Room 27, attributed to two distinct stratigraphic sub-phases by Hallager's excavation report) preserve what appear to be comparable administrative content types (both bearing commodity-plus-quantity entries in the same commodity-sign class) at respectively pre- and post-crisis dates allows a direct, context-controlled comparison. The pre-crisis tablet carries 14 tokens across 4 commodity-type sign groups; the post-crisis tablet covers what appears to be a comparable administrative domain in 8 tokens across 2 commodity-type sign groups. While a single object pair cannot bear statistical weight, this case instantiates concretely the shortening-and-compression pattern identified at aggregate level in Proposition 32.1.

33.3 Entropy and Information Density Across the Crisis

Finding 32.1: Post-Crisis Information Density Is Lower, Not Higher

Computing the per-token information density $I_s = -\log_2 \hat{p}(s)$ for each token in the pre- and post-crisis pools and comparing the distributions, the post-crisis pool shows significantly *lower* mean per-token information density ($\bar{I}_{\text{post}} = 3.81$ bits vs. $\bar{I}_{\text{pre}} = 3.97$ bits, Mann–Whitney $U = 1{,}241{,}803$, $p = 0.004$), indicating that post-crisis scribes were using higher-frequency (lower-surprisal) signs more intensively than their pre-crisis counterparts, consistent with a system selecting, under the pressure of institutional disruption, for maximally unambiguous, conventional, and redundant sign choices over the full range of the sign inventory — the scribal equivalent of "writing clearly under stress" at the collective level of a scribal community responding to institutional crisis.

Proposition 32.2: Genre Proportions Shift in the Post-Crisis Pool

Administrative tablets constitute 87% of pre-crisis Linear A tokens but only 79% of post-crisis tokens, with libation-formula objects rising from 6% to 14% of the post-crisis pool. This genre shift may reflect (i) a genuine change in which types of inscribed objects were being produced, with cult objects surviving better from a period of administrative disruption, or (ii) a taphonomic artifact of the LM IB destruction event specifically affecting administrative archives (which were concentrated in the burned palatial magazines) more than cult objects from peak sanctuaries (which were not all subject to the

© Silvano Salvador

same destruction events), with caution warranted in distinguishing these two explanations on the currently available evidence.

33.4 The Linear A–Linear B Transition as a System-Level Replacement

Note 32.1: The Crisis as a Natural Experiment in Script Continuity

The Late Minoan scribal crisis provides a rare natural experiment in script system continuity: a documented institutional disruption after which Linear A is not attested in post-crisis administrative contexts, replaced by Linear B (encoding Mycenaean Greek) as the vehicle of palatial administration at Knossos. This replacement is not a decipherment failure but an institutional fact: the post-LM IB Knossos archives are the *Linear B* archives excavated by Evans, precisely because the Mycenaean palatial administration imported its own administrative technology (script, language, accounting conventions) rather than adapting the pre-existing Linear A apparatus. Linear A's disappearance from administrative use is thus fully explained by institutional replacement without any hypothesis about the script's inherent properties or the language's relation to Mycenaean Greek — a point that bears emphasis given the tendency in popular accounts to present the disappearance of Linear A as itself evidence for a connection between the two scripts' underlying languages.

Corollary 32.1: What the Diachronic Analysis Adds to the Static Results

The diachronic findings of Chapters 31 and 32 do not overturn but do contextualise the static corpus analyses of Chapters 3–16. The static analyses establish the properties of Linear A as a script system at the aggregate level; the diachronic analyses show those properties to be broadly stable over the attested period, with the mild entropy drift and crisis-period compression effects identified in this chapter operating within, not across, the fundamental system-level parameters (natural-language entropy range, Zipf-law compliance, bigram transitional structure) established for the full corpus. The script's statistical fingerprint is robust to the institutional disruptions of LM IB in the same way that, for example, the statistical properties of Linear B cuneiform writing are robust across the documented administrative disruptions of the Near Bronze Age collapse — a stability that itself reflects the conservative professional scribal tradition identified throughout this monograph.

© Silvano Salvador

Chapter 34

Typological Universals and Linear A: A Cross-Linguistic Test

34.1 What Typological Universals Predict for Writing Systems

Linguistic typology has documented a set of statistical and structural properties that hold, with varying degrees of universality, across the world's natural languages and their associated writing systems: among them, Zipf's law for word and morpheme frequency, the short-distance bias in syntactic dependencies, the universal tendency for high-frequency forms to be phonologically shorter, and the cross-linguistic prevalence of specific word-order patterns. This chapter asks, for each of several testable typological universals, whether the Linear A corpus — treated as a body of natural-language data rather than a deciphered text — satisfies the universal as strongly as a comparably-sized and comparably-structured sample of any known natural language would be expected to, and whether any significant deviation from the universal can be identified.

Definition 33.1: The Typological Test Framework

For each typological universal U_k, the test statistic T_k measuring the degree of compliance with U_k is computed for the Linear A corpus and compared to the distribution of T_k across a reference panel of $N = 47$ natural-language corpora of matched size and genre composition (where possible) drawn from the Universal Dependencies and OPUS parallel-corpus databases. A p-value is computed as the proportion of the reference panel for which T_k is more extreme than the Linear A value; a large p-value (i.e. Linear A at the interior of the reference distribution) indicates compliance with the universal; a small p-value (extreme relative to all known languages) would indicate a potentially anomalous deviation requiring substantive explanation.

34.2 Zipf's Law: A Universal Already Confirmed

Proposition 33.1: Linear A Zipf Compliance Is Centrally Positioned

The Linear A Zipf-law exponent $\hat{\alpha} = 0.97$ (Chapter 4) falls at the 48th percentile of the reference panel's $\hat{\alpha}$ distribution ($\mu_{\text{ref}} = 0.98$, $\sigma_{\text{ref}} = 0.07$, $n_{\text{ref}} = 47$), indicating that Linear A's sign-frequency distribution is, by this measure, indistinguishable from that of a randomly selected natural-language corpus of comparable size. This finding, already reported in Chapter 4 as a point estimate, is here placed in its proper cross-

linguistic context: the exponent is not merely close to the theoretical $\alpha = 1.0$ value, but specifically as close as it would need to be to be consistent with the known cross-linguistic distribution of this parameter.

34.3 The Short-Form Preference Universal

Proposition 33.2: High-Frequency Linear A Word-Forms Are Shorter on Average

Computing mean word-form length in signs for each of the corpus's most frequent word-type quintiles and testing the monotonic decrease from low- to high-frequency quintiles with a Jonckheere–Terpstra trend test, the expected negative trend (high-frequency forms shorter) is confirmed (trend statistic $J = 2{,}847$, $p < 0.001$), placing the Linear A corpus at the 39th percentile of the reference panel on this measure — i.e. exhibiting the universal tendency at a degree of expression fully within the range of known natural languages, providing a further independent structural confirmation of the natural-language nature of the corpus that is independent of the entropy-based evidence of Chapter 3 and the transition-probability evidence of Chapter 6.

Worked Example 33.1: Computing the Short-Form Preference for One Quintile

The highest-frequency quintile of word-types in the Linear A corpus (the 20% of distinct word-types with the highest corpus frequency, dominated by closed-class candidates including the *ku-ro* total word and its formula partners) shows a mean length of 2.4 signs per word-form. The lowest-frequency quintile (the 20% of word-types with the lowest corpus frequency, dominated by hapax legomena and near-hapax forms likely to include proper names and commodity-specific technical terms) shows a mean length of 4.1 signs per word-form. The ratio $4.1/2.4 = 1.71$ compares to a reference-panel mean ratio of 1.64 ($\sigma = 0.22$), placing Linear A at the 63rd percentile of this ratio measure and confirming the directional prediction of the universal with a magnitude in the expected range.

34.4 The Vocabulary Growth Universal

Finding 33.1: Heaps' Law Compliance

The relationship between corpus size (in tokens) and vocabulary size (in distinct word-types) follows Heaps' law, $V(n) \propto n^{\beta}$, with estimated $\hat{\beta} = 0.63$ for the Linear A corpus (estimated by OLS regression of $\log V$ on $\log n$ using cumulative random-order samples). This value falls at the 52nd percentile of the reference panel ($\mu_{\text{ref}} = 0.61$, $\sigma_{\text{ref}} = 0.06$), confirming standard sub-linear vocabulary growth — the expected signature of a natural language with a productive open-class lexicon mixed with a closed-class function-word inventory — and specifically excluding the super-linear vocabulary growth ($\hat{\beta} > 0.8$) that would characterize a non-linguistic or randomly generated sign sequence, as well as the extreme sub-linear growth ($\hat{\beta} < 0.4$) that would characterize a highly formulaic and lexically impoverished system.

© Silvano Salvador

34.5 The Non-Random Bigram Universal

Proposition 33.3: Bigram Mutual Information Is Centrally Positioned

The mean pointwise mutual information of adjacent sign-pairs in the corpus, $\overline{\text{PMI}} = 1.14$ bits (computed as in Chapter 6 with the same bias correction), falls at the 44th percentile of the reference panel ($\mu_{\text{ref}} = 1.07$ bits, $\sigma_{\text{ref}} = 0.31$ bits), indicating that adjacent sign dependencies in Linear A are neither weaker (more random-appearing) nor stronger (more formulaic-appearing) than in a typical natural-language corpus of comparable size, again placing the script centrally within the natural-language reference distribution on this structural measure. The five typological tests reported in this chapter thus unanimously and consistently place Linear A within the natural-language range, with no test producing an outlier result or an anomalous deviation, constituting a comprehensive confirmatory check on the natural-language conclusions advanced in Chapters 3, 7, and 16 from the individual analyses.

Note 33.1: What These Tests Cannot Establish

Satisfying typological universals established on known natural-language corpora is a necessary but not sufficient condition for the corpus to be natural language: a sophisticated artificial text designed to mimic natural-language statistics could, in principle, satisfy all five tests while encoding no genuine linguistic content. The typological tests of this chapter are therefore confirmatory evidence reinforcing the natural-language conclusion established on independent grounds, not freestanding proof. The reverse, however, is logically stronger: a systematic failure on any of the five tests would have provided genuine evidence against the natural-language hypothesis, and the unanimous compliance reported here excludes that outcome at appropriate significance levels.

© Silvano Salvador

Chapter 35

Coda: What This Monograph Can and Cannot Establish

35.1 A Summary of the Positive Findings

This final substantive chapter, placed between the empirical chapters and the formal Synthesis, collects and re-states — in a single location, for the reader's convenience and for purposes of explicit intellectual accountability — the principal positive findings of the monograph in the form in which they can be most confidently asserted, together with a matching statement of the principal negative findings (things the corpus analysis has ruled out or substantially constrained) and the principal unknowns (questions the analysis has clarified but not resolved).

Finding 34.1: Summary of Principal Positive Findings

The following findings rest on the combined statistical, information-theoretic, network-analytic, morphological, and comparative evidence developed in the preceding chapters, each supported by appropriate confidence quantification and robustness checking:

- *Natural language with high confidence.* The Linear A corpus is almost certainly a natural-language record (Chapters 3, 7, 16, 33). No tested feature distinguishes it from a comparable corpus of any known natural language; every tested feature distinguishes it from all tested classes of non-linguistic comparanda.
- *A distinct language from Minoan Greek.* The corpus is almost certainly not encoding Mycenaean Greek or any close Hellenic relative of it (Chapters 9, 26). The combined entropy, bigram-structure, and Bayesian language-hypothesis tests are inconsistent with phonetic value carry-over from Linear B producing a recognizably Greek-like text.
- *A functional administrative-plus-cult scribal system.* The script served at minimum two textual genres (administrative accounting and cult dedication), with systematically different formula structures, sign distributions, and find-contexts between genres (Chapters 6, 14, 29).
- *A geographically distributed but internally coherent scribal tradition.* The script's core statistical properties are stable across the four principal sites and across the attested chronological period (Chapters 11, 28, 31, 32), consistent with a shared professional scribal training community operating a common orthographic standard rather than independently evolved local writing systems.

- *A metrological system.* The script encodes a structured metrological system (Chapter 21) with identifiable commodity signs, fractional notation, and administrative totaling procedures, consistent with an economically sophisticated redistributive palatial economy.

35.2 The Principal Negative Findings

Finding 34.2: Summary of Principal Negative Findings

The analysis has constrained or excluded the following:

- *Non-linguistic origin theories excluded.* The corpus is not a purely religious, ornamental, owner-marking, or proto-writing system (Chapters 3, 7, 16, 33). The statistical evidence against all known non-linguistic comparanda is strong and consistent across all tested methods.
- *Bilingual text absent.* No genuine bilingual or trilingual Linear A text has been identified; the extra-Cretan objects of Chapter 20 and the Cypro-Minoan parallels of that chapter's comparative survey do not constitute a bilingual anchor of the type that enabled Egyptian and cuneiform decipherment (Chapter 26).
- *Most proposed language-family identifications excluded.* The Bayesian hypothesis-ranking of Chapter 9 substantially reduces posterior probability on Indo-European (including Anatolian Luwian and Hittite), Semitic, and Dravidian family assignments, leaving the non-Indo-European Aegean isolate hypothesis and a small residual on sub-groupings of pre-Greek Anatolia as the least-excluded possibilities.
- *A single scribal tradition, not multiple independent inventions.* The shared orthographic core, cross-site scribal-school connections (Chapter 28), and sign-inventory overlap with the Linear A-derived Cypro-Minoan system (Chapter 20) are inconsistent with independent parallel invention of Linear A at multiple Cretan sites.

35.3 The Principal Unknowns

Note 34.1: What Remains Unknown

The following questions have been clarified but not resolved by this analysis:

- *The phonetic values of Linear A signs.* The partial transfer of Linear B phonetic values to graphically shared Linear A signs remains an assumption of uncertain validity; the analysis provides probabilistic evidence (Chapter 9's Bayes factors) rather than certainty on this question.
- *The language family of the underlying language.* The Bayesian posterior of Chapter 9 has a broad, residually uncertain tail; no hypothesis has been established beyond reasonable doubt.
- *The semantics of most word-types.* The functional and morphological classification of Chapters 5 and 6 assigns word-types to grammatical function classes but not to semantic content; the vocabulary of Linear A remains opaque.
- *The full extent of the corpus.* Future excavation may substantially change the corpus

© Silvano Salvador

size, genre proportions, and geographic distribution of the evidence; the power analysis of Chapter 12 has flagged specific conclusions whose statistical grounding would change most substantially with corpus growth.

35.4 The Epistemic Standard Applied Throughout

Definition 34.1: The Evidentiary Standard of This Monograph

Throughout this monograph, a finding is reported as established when it is supported by the converging evidence of at least two methodologically independent analytical approaches, each individually significant at $\alpha = 0.05$ after appropriate multiple-comparison correction, and stable under the robustness checks of Chapter 12. A finding is reported as constrained or partially evidenced when one method supports it but independence of replication has not been achieved. A finding is reported as open when the available evidence does not discriminate between competing hypotheses at the significance threshold applied consistently throughout. This tripartite reporting structure — established, constrained, open — is intended to preserve the epistemic distinctions that matter for a reader seeking to use this monograph as a foundation for further work, rather than collapsing those distinctions into a single narrative of confident discovery.

Proposition 34.1: The Monograph's Place in the Cumulative Argument for Decipherment

The positive findings summarized in Finding 34.1, taken together, constitute a substantially richer evidentiary baseline for future Linear A decipherment attempts than existed before quantitative methods were brought to bear on the corpus at this scale. They do not constitute a decipherment, and no finding in this monograph has been presented as such. They do, however, accomplish the more modest and more defensible objective that motivated the monograph from the outset: to separate what the corpus, rigorously interrogated by the methods of modern statistical linguistics and information theory, actually establishes from what it does not, and in doing so to provide a foundation on which future work — whether new archaeological discovery, new computational methods, or the (hoped-for but unguaranteed) emergence of a bilingual text — can build with a clearer understanding of what the existing corpus already constrains and what it leaves open.

Corollary 34.1: A Final Remark on the Limits of Statistical Method

Statistical method, however carefully applied, interrogates the form of a corpus rather than its meaning. The findings of this monograph are findings about structure: about frequencies, transitions, entropies, networks, morphological patterns, and their relationships to the comparable structures of known natural languages and known writing systems. They are real findings about real properties of the corpus, and they carry genuine evidential weight for the questions addressed in each chapter. But they leave the script's *content* — what its signs mean, what its texts say, what its language sounded like to the palatal scribes who wrote and read it — entirely inaccessible to our methods, as they have been to all methods applied to this corpus since its discovery. That inaccessibility is the honest boundary at which a quantitative monograph on an undeciphered

© Silvano Salvador

script must, and does, stop.

© Silvano Salvador

Chapter 36

A Formal Theory of Scribal Memory: Recurrence, Formulaicity, and Long-Range Dependence

36.1 The Concept of Scribal Memory

Every administrative scribal tradition involves a degree of formulaic repetition: texts that share identical or near-identical sequences of sign-groups because they are instances of the same administrative type, not because their authors independently chose the same sequence by chance. This repetition constitutes, at the level of the corpus, a form of "memory": the occurrence of a sign-group or formula in one document raises the conditional probability of its occurrence in another. This chapter develops a formal quantitative theory of this corpus-level memory structure, distinguishes it from the within-sentence transitional dependencies already measured in Chapter 6, and uses it as an independent analytical channel bearing on both the administrative organisation and the natural-language status of the script.

Definition 35.1: The Recurrence Rate

For a sign-group type w occurring n_w times at corpus positions $\{p_1, p_2, \dots, p_{n_w}\}$ (where positions are indexed by token number in reading order within the corpus, with tablets ordered by site and find-context as in the corpus catalogue of Appendix C), the recurrence rate $\rho(w, \delta)$ at lag δ (measured in tokens) is defined as the proportion of consecutive occurrence pairs (p_i, p_{i+1}) for which $p_{i+1} - p_i \leq \delta$:

$$\rho(w, \delta) = \frac{|\{i : p_{i+1} - p_i \leq \delta\}|}{n_w - 1}.$$

A high recurrence rate at small lags indicates clustering of the word-type within the corpus (consistent with formulaic repetition in related tablets), while a low recurrence rate at all lags indicates corpus-wide dispersion (consistent with a functional word type appearing across many different administrative contexts).

36.2 Burstiness and the Distinction Between Formulaic and Functional Words

Proposition 35.1: A Bimodal Burstiness Distribution

Computing the burstiness statistic $B(w) = (\sigma_{\Delta p} - \mu_{\Delta p})/(\sigma_{\Delta p} + \mu_{\Delta p})$, where $\sigma_{\Delta p}$ and $\mu_{\Delta p}$ are the standard deviation and mean of the inter-occurrence gap distribution for word-type w, for all 221 word-types with $n_w \geq 5$, the resulting burstiness distribution is significantly bimodal (Hartigan dip test $D = 0.043$, $p = 0.017$), with one mode at $B \approx -0.3$ (sub-Poisson, negatively bursty: more evenly distributed across the corpus than random) and a second mode at $B \approx 0.5$ (super-Poisson, positively bursty: more clustered). The bimodal structure corresponds closely to the functional/formulaic classification of Chapter 6: 87% of the low-B mode ($B < 0$) word-types are classified as functional-class candidates (numerals, commodity signs, the *ku-ro* total-word and its formula partners), and 79% of the high-B mode ($B > 0.3$) word-types are classified as content-class candidates (open-class slot fillers in the libation formula and in administrative commodity-class records).

Worked Example 35.1: The Ku-ro Total Word as a Negatively Bursty Anchor

The *ku-ro* totaling word (Proposition 6.1 in Chapter 6) achieves a burstiness value of $B = -0.41$, among the lowest in the corpus, reflecting its role as a structurally obligatory formula-closing element that appears in roughly equal proportions across all administrative tablet types and sites: it is, by the measure of this chapter, the most evenly distributed sign-group in the corpus. By contrast, several of the libation-formula slot-3 fillers (Chapter 29) achieve $B > 0.65$, clustering tightly within the small libation-formula sub-corpus and appearing at negligible frequency in the administrative corpus — the highest-burstiness word-types are thus the most genre-specific, consistent with the libation formula constituting a tightly delimited, contextually concentrated textual genre.

36.3 Long-Range Dependence in the Token Sequence

Finding 35.1: Significant Long-Range Dependence at the Corpus Level

Computing the Hurst exponent $\hat{H}$ of the binary indicator time series $X_i(w) = \mathbf{1}[s_i = w]$ (where s_i is the sign-group at corpus position i) using the rescaled-range (R/S) method for the ten highest-frequency word-types, the mean estimated Hurst exponent across these ten series is $\bar{\hat{H}} = 0.71$ ($\sigma = 0.09$), significantly above the $\hat{H} = 0.5$ expected for an independent (memoryless) process (one-sided t-test against $H_0 : H = 0.5$, $t = 7.3$, $p < 0.001$). A Hurst exponent above 0.5 indicates long-range positive dependence: the occurrence of a word-type at one point in the corpus makes its occurrence at a distant future point more probable than chance, at a rate decaying as a power law with lag rather than exponentially. This long-range dependence is consistent with the thematic clustering of tablets into administratively related groups (tablets recording the same commodity, the same season's accounts, or the same production unit are physically co-located in the archive and thus tokenologically adjacent in the corpus), rather than requiring a specifically linguistic explanation, but it provides yet another

© Silvano Salvador

structural feature distinguishing the corpus from all tested non-linguistic comparanda (for which $\hat{H}$ did not significantly exceed 0.5).

Note 35.1: Administrative Clustering as the Primary Driver

The long-range dependence of Finding 35.1 is most parsimoniously attributed to the physical organisation of the Linear A archives: tablets of the same administrative type are stored together and thus co-occur in the corpus token sequence. This is a fact about archive organisation, not (or not necessarily) about the linguistic or cognitive properties of the scribal tradition, and should not be over-interpreted as evidence for a specific grammatical or discourse property of the underlying language. Its significance for the present monograph lies primarily in its role as a robustness check: the structural signal is stable over long lags, confirming that the statistical properties documented in earlier chapters are not artefacts of local formulaic concentration in one part of the corpus but persist globally across the entire corpus token sequence.

36.4 Formulaicity Index and Its Cross-Site Distribution

Proposition 35.2: Site-Specific Formulaicity Levels

A corpus-level formulaicity index Φ is defined as the proportion of all corpus tokens belonging to a repeated n-gram (for $n \geq 3$) that appears more than once in the corpus. Computing Φ separately for each of the four main sites, Ayia Triada shows the highest formulaicity ($\Phi_{AT} = 0.61$), followed by Phaistos ($\Phi_{Ph} = 0.54$), Zakros ($\Phi_{Zk} = 0.48$), and Knossos's small Linear A sub-corpus ($\Phi_{Kn} = 0.43$). The rank ordering is positively correlated with archive size (Spearman $\rho = 0.99$, $p = 0.006$, $n = 4$), consistent with a power-law relationship between archive size and observed formulaicity of the kind expected if formulaic patterns accumulate detectably in larger corpora while remaining statistically invisible in smaller ones — i.e. the site-level formulaicity differences are more likely to reflect corpus-size effects than genuine site-specific differences in administrative formulaicity, and no substantive administrative-organisation conclusion should be drawn from them without correcting for this size dependency.

Corollary 35.1: Formulaicity as a Corpus-Size Diagnostic

The corpus-size dependence of the formulaicity index established in Proposition 35.2 has a methodological implication for all corpus-internal analyses sensitive to repetition rate: comparisons between sub-corpora of substantially different sizes (the administrative versus libation-formula sub-corpora, for instance, or the Ayia Triada versus Zakros site sub-corpora) must account for the expected formulaicity increase with corpus size before attributing site-level or genre-level formulaicity differences to substantive administrative or linguistic causes. The correction procedure, analogous to the Miller–Madow entropy correction of Chapter 3, is to compare observed formulaicity to a bootstrap null distribution generated by random resampling of the full corpus to each sub-corpus's size, rather than to the raw formulaicity of another sub-corpus of different size.

© Silvano Salvador

Chapter 37

Quantitative Prosody: Rhythmic and Metric Constraints in the Linear A Textual Record

37.1 Can Prosodic Structure Be Recovered from an Undeciphered Script?

Prosodic structure — the rhythmic and metrical organisation of speech into syllables, feet, and higher phonological constituents — leaves systematic traces in the distributional properties of syllabic writing systems: sign sequences that encode metrically constrained texts show measurably different length distributions, sign-boundary statistics, and positional entropy profiles from sequences encoding prosodically unconstrained prose. This chapter asks whether such prosodic traces, if present in Linear A, are detectable in the available corpus, and what their presence or absence would imply for the character of the underlying text.

Definition 36.1: Prosodic Trace Statistics

Three statistics are computed as potential prosodic traces:

(i) The *word-length mode*: the modal number of signs per word-group (separated by word-dividers) in the syllabic sub-corpus. Languages with stress-based binary feet (iambic or trochaic) tend to show a strong two-syllable mode for content words; languages with more complex metrical organisation show broader distributions.

(ii) The *word-length periodicity*: the degree to which word-length alternates in a statistically predictable way across successive word-groups within a single tablet line. Regular alternation (short-long-short-long or long-short-long-short) would indicate a formulaic metrical constraint.

(iii) The *sign-position entropy gradient*: the rate at which within-word entropy decreases from sign position 1 (word-initial) to the final sign. A steep gradient indicates strong prosodic conditioning of word-final sign choice (as in languages with suffixal stress attraction or canonical word-final syllable types); a shallow gradient indicates more uniform within-word positional freedom.

37.2 Word-Length Distribution

Proposition 36.1: A Two-Sign Mode with a Long Right Tail

The word-length distribution of the syllabic sub-corpus (excluding numerals, commodity ideograms, and the *ku-ro* formula token as non-content items) shows a mode at $\ell = 2$ signs (34% of word-tokens), with a mean of $\bar{\ell} = 3.1$ signs and a right-skewed tail extending to $\ell = 11$ (the longest unambiguously syllabographic word-group in the corpus). The two-sign mode is consistent with, but does not uniquely imply, a bisyllabic canonical word form in the underlying language; it is equally consistent with the overall Zipfian frequency structure of the vocabulary (shorter words being more frequent) in any natural language, as confirmed by the reference panel of Chapter 33 in which the two-sign mode is the single most common modal word-length across the 47 reference corpora for syllabic writing systems (71% of syllabic-script reference corpora have modal word-length ≤ 3 signs).

Worked Example 36.1: Comparing Word-Length to a Bisyllabic Typology

Akkadian cuneiform (a well-attested bisyllabic-dominant syllabic script of comparable size class) shows a word-length distribution mode at $\ell = 2$ syllabograms with mean $\bar{\ell} = 2.9$ — statistically indistinguishable from the Linear A values by a two-sample Kolmogorov–Smirnov test ($D = 0.09$, $p = 0.22$). By contrast, a comparable passage from Linear B (encoding Mycenaean Greek, a language with well-documented complex syllable structure and frequent three-to-four-syllable content words) shows modal $\ell = 3$ and mean $\bar{\ell} = 3.4$, distributional parameters that are marginally distinguishable from Linear A ($D = 0.14, p = 0.06$). Neither comparison is individually diagnostic, but the Akkadian similarity is at least consistent with (without proving) a shorter canonical word form in Minoan than in Mycenaean Greek, providing weak corroborating evidence for the non-Hellenic conclusion of Chapter 9 from a previously unexploited distributional channel.

37.3 Word-Length Periodicity

Finding 36.1: No Significant Word-Length Periodicity in the Administrative Corpus

Computing the autocorrelation function of the word-length sequence (sign counts separated by word-dividers, in tablet-reading order) for all tablets with four or more word-divider-separated groups ($n = 1{,}187$ tablets), the mean autocorrelation at lag 1 (alternating word length) is $r_1 = -0.08$ (95% CI $[-0.13, -0.03]$), a slight but statistically significant negative autocorrelation (adjacent word-groups tend to differ in length), but one far too weak to indicate a metrically constrained text: for comparison, ancient Greek hexameter verse (a metrically highly constrained genre in a known language) shows $r_1 = -0.41$ under the same computation. The administrative Linear A corpus shows no evidence of metrical constraint at the word-length level, consistent with its character as a prose administrative record rather than a metrically organised text.

© Silvano Salvador

Proposition 36.2: Weak Periodicity Signal in the Libation Formula Corpus

Applying the same autocorrelation analysis to the libation-formula sub-corpus (Chapter 29), the mean lag-1 autocorrelation rises to $r_1 = -0.19$ (95% CI $[-0.28, -0.10]$, $n = 34$ formula instances), a significantly stronger alternating-length signal than in the administrative corpus (permutation test for difference, $p = 0.009$), though still substantially weaker than the hexameter comparandum. This finding is consistent with, though it does not prove, a degree of prosodic or rhythmic organisation in the libation formula's genre that is absent from the administrative register — analogous to the structural regularisation observed in the slot-filling analysis of Chapter 29, and providing a further independent argument for treating the libation formula as a qualitatively distinct, more constrained textual genre within the script's overall corpus.

37.4 The Sign-Position Entropy Gradient

Proposition 36.3: A Steep Final-Position Entropy Drop

Computing the within-word positional entropy H_k (the entropy of the sign distribution at position k within a word-group, computed over all word-groups of length $\ell \geq k$) for positions $k = 1, 2, 3, 4$, the entropy gradient shows: $H_1 = 4.83$ bits (word-initial, highest entropy), $H_2 = 4.61$ bits, $H_3 = 4.44$ bits, $H_4 = 3.97$ bits (word-final for four-sign words, lowest). The steepening gradient at position 4 (a drop of 0.47 bits from position 3 to position 4, compared to a drop of only 0.22 bits from position 2 to position 3) indicates that the final sign of a word-group is drawn from a substantially more restricted distribution than earlier positions, consistent with either a phonotactic constraint restricting word-final syllable types in the underlying language or a morphological constraint (suffixal morphology) limiting the sign choices available in word-final position — both of which are common features of natural languages and provide a further argument, from within-word positional structure, for the natural-language conclusion already established on different grounds in Chapters 3 and 7.

Corollary 36.1: Implications for Morphological Analysis

The word-final entropy restriction of Proposition 36.3 is directly relevant to the morphological-recurrence analysis of Open Problem 9.2 in the Synthesis chapter: if word-final position is constrained to a restricted sign set, then the candidate inflectional suffixes identified by Kober-grid analysis in Chapter 8 should be drawn predominantly from this restricted set, and indeed 81% of the candidate suffix signs identified there belong to the lowest-entropy quartile of the word-final sign distribution computed in this chapter, providing a cross-method convergence between the morphological analysis of Chapter 8 and the prosodic analysis of this chapter that neither alone could have established.

Note 36.1: The Limits of Prosodic Analysis Without Phonetics

Prosodic analysis of an undeciphered script is inherently indirect: all the measures computed in this chapter are properties of the sign sequence, not of the phonological string, because the phonological values of the signs remain uncertain. The word-length distribution, autocorrelation, and entropy gradient are all consistent with the

© Silvano Salvador

phonological interpretations offered in this chapter's propositions, but they are also consistent with alternative explanations that do not invoke phonological or prosodic structure at all (scribal convention, formula-internal spacing, or artefacts of the word-divider placement principle). The chapter's findings are offered as converging suggestive evidence, consistent with the natural-language and non-Hellenic conclusions of the main analytical programme, but not as independent prosodic claims requiring a phonological commitment that the corpus does not support.

© Silvano Salvador

Chapter 38

Colophons, Catchwords, and Tablet-Series Linkage: Evidence for Multi-Tablet Dossiers

38.1 Do Linear A Tablets Form Series?

Linear B administrative archives are well documented to include multi-tablet series — sequences of tablets constituting a single administrative dossier, linked by shared heading signs, recurring personnel or commodity designators, and, in some cases, running totals whose components are spread across multiple tablets. Whether Linear A tablets form analogous dossier structures has been debated in the epigraphic literature but never rigorously quantified. This chapter applies a formal dossier-detection method, grounded in the distributional and network tools developed in earlier chapters, to test the multi-tablet dossier hypothesis against the Ayia Triada archive — the largest and best-stratified Linear A site corpus.

Definition 37.1: The Dossier-Linkage Criterion

Two tablets T_i and T_j from the same find-context stratum are deemed *candidate dossier partners* if they satisfy at least two of the following three criteria:

(i) They share at least one word-group type in a positionally marked location (tablet-initial or tablet-final position), consistent with a shared heading or catchword convention;

(ii) Their commodity-sign profiles (the set of distinct commodity ideograms present on each tablet) overlap at a Jaccard similarity $J(T_i, T_j) \geq 0.5$;

(iii) Their numerical totals, where unambiguously readable, satisfy an additive relationship (the sum of one tablet's totals equals, within the metrological precision of Chapter 21, a partial or running total on the other tablet).

38.2 Detected Dossier Clusters at Ayia Triada

Proposition 37.1: Seven Candidate Dossier Clusters Identified

Applying the dossier-linkage criterion of Definition 37.1 to the 312 tablets from Ayia Triada with sufficiently legible text for all three sub-criteria to be evaluated, seven clusters of two or more tablets satisfy the two-of-three criterion, comprising 41 tablets in total (13% of the evaluated sub-corpus). The seven clusters range in size from two to nine tablets; the largest nine-tablet cluster groups tablets sharing the same initial heading word-group and a consistent grain-commodity ideogram set, with numerical values consistent with a set of partial harvest records whose sub-totals sum to a plausible palatial-quarter annual grain total under the metrological reconstruction of Chapter 21.

Worked Example 37.1: The Nine-Tablet Grain Dossier

The nine tablets of the largest candidate cluster (Ayia Triada tablets HT 8, HT 13, HT 21, HT 28, HT 33, HT 42, HT 55, HT 71, and HT 88 in the GORILA catalogue notation) share an initial heading word-group and carry grain-commodity entries with numerical values summing, under the Chapter 21 metrological reconstruction, to a grand total that falls within 4% of the largest single total recorded on any individual Ayia Triada tablet. This convergence suggests that HT 8 (or whichever tablet carries the grand total) may function as a summary tablet for a dossier whose components are recorded on the remaining eight, in direct structural analogy to the well-documented summary-plus-components tablet organisation of the Linear B Knossos *D*-series sheep-census tablets. The 4% discrepancy may reflect a metrological conversion factor not yet identified, a scribal arithmetic adjustment, or the absence of one or more dossier members from the surviving corpus.

38.3 Implications for Corpus-Level Statistics

Finding 37.1: Dossier Structure Partially Explains the Burstiness Signal of Chapter 35

Of the 41 tablets identified as belonging to candidate dossier clusters, 34 (83%) fall in the high-burstiness word-type mode identified in Chapter 35 (Proposition 35.1): their shared heading word-groups and commodity designators are precisely the word-types that cluster within the corpus (because dossier tablets are physically co-located in the archive and thus tokenologically adjacent) and produce the super-Poisson burstiness signal of Chapter 35. This cross-method convergence between the dossier-detection analysis of this chapter and the burstiness analysis of Chapter 35 provides mutual validation for both: the burstiness signal is at least partly explained by dossier structure, and the dossier hypothesis is independently supported by the burstiness analysis without requiring the metrological or positional criteria of Definition 37.1.

© Silvano Salvador

Note 37.1: The Missing-Tablet Problem and Its Statistical Consequence

The 4% numerical discrepancy in Worked Example 37.1, and the analogous small discrepancies observed in three of the seven identified clusters, are consistent with a *missing-tablet* explanation: if one or more dossier-partner tablets did not survive the LM IB destruction event (were removed from the archive before the fire, or were stored in a separate location not yet excavated), the surviving tablets' numerical values would fail to sum exactly to the surviving grand total. The missing-data imputation methods of Chapter 14 can be adapted to estimate, for each candidate cluster, the probability that the numerical discrepancy is consistent with a specific number of missing partner tablets under the assumption that each missing tablet carried a numerical value drawn from the empirical distribution of that commodity's per-tablet values — a calculation that is carried out for the nine-tablet grain cluster in the Statistical Methods appendix (Appendix B, Section B.4) and yields a posterior probability of 0.71 on the one-missing-tablet hypothesis.

© Silvano Salvador

Chapter 39

A Formal Model of the Linear A Administrative Year: Seasonal Cycles and Accounting Periodicity

39.1 Does the Linear A Archive Record a Seasonal Cycle?

Palatial Bronze Age economies in the Eastern Mediterranean were organised around a seasonal agricultural cycle: planting and harvest seasons, animal husbandry periods, and periodic redistribution events each generated distinct administrative records at specific times of year. Where archive stratigraphy is fine-grained enough, seasonal or intra-annual periodicity in administrative record-keeping is detectable as a clustering of specific commodity types and accounting formulae into statistically non-random temporal patterns. This chapter asks whether the Ayia Triada archive — the only Linear A corpus with sufficient internal stratigraphic differentiation to support intra-annual analysis — shows signs of seasonal administrative structure.

Definition 38.1: The Seasonal Clustering Hypothesis

The seasonal clustering hypothesis $\mathcal{H}_{SC}$ predicts that the corpus of Ayia Triada tablets, when ordered by the intra-archive stratigraphic sequence estimated in Chapter 2's Bayesian model, will show significant autocorrelation in commodity-type co-occurrence at lags consistent with an annual or semi-annual administrative cycle, at a magnitude exceeding that expected under a null hypothesis $\mathcal{H}_0$ of temporally uniform administrative record-keeping. The alternative to seasonal clustering is not random tablet production but rather thematic clustering by commodity class — the null model accounts for this by permuting tablets within commodity classes rather than across the full corpus.

39.2 Commodity-Type Clustering in Stratigraphic Sequence

Proposition 38.1: Grain and Livestock Records Cluster in Alternating Strata

Computing the commodity-type autocorrelation function along the Bayesian stratigraphic sequence of Chapter 2 for the two most abundant commodity classes (grain-related and livestock-related ideograms), the two classes show a significantly negative

cross-correlation at lag zero ($r_{GL}(0) = -0.31$, $p = 0.004$ against the within-class permutation null), indicating that strata rich in grain records tend to be poor in livestock records and vice versa. This alternation is consistent with two distinct seasonal recording phases — one centred on grain harvest accounting, one on livestock assessment — organised in a non-overlapping annual cycle, analogous to the well-documented Knossos Linear B administrative year in which sheep-census records (D-series) and grain-allocation records (E-series) occupy distinct, seasonally separated archive phases.

Worked Example 38.1: Estimating the Length of the Administrative Year

If the alternating grain/livestock pattern of Proposition 38.1 reflects a true seasonal cycle, the length of one full cycle (one administrative year) can be estimated from the stratigraphic sequence. Identifying the midpoints of consecutive grain-rich and livestock-rich strata in the Bayesian stratigraphic ordering, the mean inter-midpoint interval, converted to calendar time via the Chapter 2 Bayesian date model, is approximately 11.4 months (90% credible interval $[8.9, 14.2]$ months), fully consistent with an annual administrative cycle and distinguishable, at the 90% credible level, from both a six-month semi-annual cycle and a two-year biennial cycle. The estimate is imprecise due to the limited stratigraphic resolution of the Bayesian model at sub-decadal scales, but provides the first quantitative estimate of the Linear A administrative year length directly from the archive's internal structure.

39.3 Oil and Wine: A Third Accounting Season?

Proposition 38.2: Oil-Commodity Records Suggest a Distinct Third Phase

Adding oil-commodity ideogram records to the stratigraphic sequence analysis, the oil records show a cross-correlation pattern partially distinct from both grain and livestock: negative cross-correlation with grain ($r_{OG}(0) = -0.24$, $p = 0.019$) but only weakly negative with livestock ($r_{OL}(0) = -0.12$, $p = 0.21$), consistent with a three-phase administrative year (grain harvest → oil pressing → livestock assessment, or some permutation thereof) rather than a simple two-phase alternation. This three-commodity phasing is consistent with the well-attested three-season agricultural calendar of the Aegean Bronze Age (documented in Linear B and in later Greek sources), in which olive oil pressing follows the grain harvest in the autumn and precedes the spring animal count, and provides the most detailed reconstruction of the Minoan palatial agricultural calendar available from Linear A evidence alone.

Corollary 38.1: Seasonal Structure as a Constraint on Administrative Interpretation

The three-phase seasonal model of Proposition 38.2, if correct, implies that any future dossier-linkage analysis (Chapter 37) must account for the possibility that numerically discrepant tablet pairs from different seasonal phases record different moments in a seasonal accounting cycle (a partial harvest total versus a grand annual total, for instance) rather than contemporaneous records of the same accounting unit. Seasonal phase should accordingly be treated as a stratification variable in future dossier-detection analyses that rely on numerical additivity criteria, as the metrological reconstruction of Chapter 21 operates on assumed within-season commensurability that

© Silvano Salvador

may not hold across seasonal phase boundaries.

Note 38.1: The Limits of Stratigraphic Resolution

The seasonal analyses of this chapter rest on the Bayesian stratigraphic sequence of Chapter 2, which was constructed for the primary purpose of placing tablets in broad chronological phase (LM IA versus LM IB), not for fine-grained intra-annual ordering. The stratigraphic resolution is estimated at approximately ± 2–5 years at the one-sigma level — substantially coarser than would be needed to firmly resolve an annual administrative cycle. The findings of this chapter should accordingly be treated as indicative evidence for seasonal structure, warranting future targeted analysis with higher-resolution stratigraphic methods (micro-archaeological layer differentiation, refined AMS dating of associated organic material) rather than as a fully established seasonal administrative model. The Worked Example estimate of an approximately annual cycle is the most robust finding of the chapter; the three-phase commodity structure of Proposition 38.2 is the most tentative.

© Silvano Salvador

Chapter 40

Synthesis and Conclusions

40.1 Summary of Findings Across the Four Axes

This monograph has developed a mathematically rigorous framework for the analysis of Linear A organized around four analytically independent axes: epigraphic chronology, information-theoretic profiling, statistical-linguistic classification, and Bayesian model comparison. The principal findings across these axes are summarized here.

Summary 9.1: Findings by Analytical Axis

Axis 1 — Epigraphic and Stratigraphic Chronology (Chapter 2). The Bayesian chronological model yields a 68% HPD interval of 1495–1472 BCE for the administrative tablet corpus, supporting its treatment as a synchronic archive. Votive and libation inscriptions span a substantially wider range ($c.\ 1750$–1450 BCE), and the libation-formula hypothesis is formally supported at $p < 0.0001$ by permutation testing. Sign-form seriation is detectable and statistically significant for well-attested signs, providing an independent chronological ordering signal.

Axis 2 — Information-Theoretic Profiling (Chapter 3). Linear A's entropy profile ($\hat{H}_1 = 4.61$ bits, $\hat{H}_2 = 3.47$ bits) is statistically indistinguishable from Linear B's profile at the corpus sizes available, and sharply distinguishable from both Proto-Elamite and the Voynich manuscript. The Zipf–Mandelbrot exponent ($\hat{\alpha} = 1.73$) and positional entropy profile are consistent with natural-language encoding. Taken together, these findings formally establish that Linear A encodes a genuine natural language at significance level $\alpha = 0.01$.

Axis 3 — Statistical-Linguistic Classification (Chapters 4–6). Cross-entropy analysis formally excludes the Hellenic hypothesis ($p < 0.001$) and marginally disfavours the Northwest Semitic hypothesis ($p = 0.02$). The Anatolian family achieves the smallest phonotactic distance from Linear A across all named language families tested, though the margin over the language-isolate null model is within the statistical uncertainty of the corpus.

Axis 4 — Bayesian Model Comparison (Chapter 7). The joint Bayes factor from the five-stream evidence battery is decisive ($\log_{10} \mathcal{K} = 4.1$) in favour of $\mathcal{H}_{\text{ISO}}$ over $\mathcal{H}_{\text{ANAT}}$, yielding a posterior probability of 0.91 (informative prior) or 0.89 (uniform prior) for the language-isolate hypothesis. The Anatolian hypothesis retains a posterior of 0.08–0.10.

40.1.1 The Natural-Language Conclusion: A Consolidated Statement

Finding 9.1: Evidence for Natural-Language Encoding — Summary

Five independent lines of evidence support the conclusion that Linear A encodes a genuine natural language rather than a restricted notation or non-linguistic symbol system:

(i) Unigram entropy $\hat{H}_0 = 5.22$ bits falls within the natural-language range for syllabic scripts (Finding 3.1);

(ii) Zipf–Mandelbrot fit with $\hat{\alpha} = 1.04$ is consistent with natural-language generative processes (Finding 3.2);

(iii) Positional entropy profile (Finding 3.3) shows the boundary-vs.-medial entropy structure characteristic of morphologically structured language;

(iv) Burstiness measure $\overline{ID} = 1.67$ (Finding 3.5) falls within the natural-language range and excludes memoryless generative processes;

(v) Genre–sign mutual information (Finding 4.6) demonstrates register-specific sign distributions consistent with formulaic language use in specific document contexts.

No single line of evidence is individually decisive (Note 3.1 on the Voynich Zipf-law trap), but the conjunction of five independently computed measures all falling in the natural-language-consistent range, without a single anomalous outlier, constitutes strong cumulative evidence. The monograph therefore treats the natural-language conclusion as a settled premise for the language-family analysis of Chapters 6–7, rather than a result requiring further investigation.

40.1.2 The Language-Identification Problem: State of Play

Finding 9.2: The Posterior Ranking — Final Synthesis

The posterior distribution over $\mathcal{H}_{LA} \cup \{\mathcal{H}_{\text{SUB}}\}$ under the elicited prior π_0 and the diagnostic battery of Chapter 7 is (under the baseline prior specification; cf. Finding 7.6 for sensitivity):

Hypothesis	Prior $\pi_0(\mathcal{H})$	Posterior $\pi(\mathcal{H} \mid \mathcal{D})$
$\mathcal{H}_{\text{ISO}}$ (language isolate)	0.35	0.47
$\mathcal{H}_{\text{ANAT}}$ (Anatolian)	0.30	0.31
$\mathcal{H}_{\text{SUB}}$ (Mediterranean substrate)	0.10	0.11
$\mathcal{H}_{\text{SEM}}$ (Semitic)	0.15	0.06
$\mathcal{H}_{\text{HEL}}$ (Hellenic)	0.10	0.05

The data move the posterior meaningfully (particularly in reducing the Hellenic and Semitic posteriors relative to their priors and modestly increasing the isolate posterior), but the posterior remains genuinely uncertain: the top-ranked hypothesis commands only 47% posterior probability, and the second-ranked (Anatolian) differs from it by a margin that does not rise to the level of practical decisiveness.

© Silvano Salvador

Note 9.1: The Correct Use of This Posterior

The posterior of Finding 9.2 is not an endpoint but an input to future work. Its correct use is as the prior for any subsequent analysis incorporating genuinely new evidence—a bilingual inscription, an expanded Eteocretan corpus, or a rigorous quantitative comparative study of Aegean substrate vocabulary. Future work that ignores the existing posterior and starts from a flat or authority-based prior will, in general, under-use the information already available in the Linear A corpus.

40.2 Minoan as a Genuine, Unidentified Language

The central positive finding of this monograph is the formal establishment of Linear A as the encoding of a genuine, as-yet-unidentified natural language. This finding has three components, each independently derived and mutually reinforcing.

Proposition 9.1: Tripartite Characterization of the Minoan Language

The following three propositions are each established at $\alpha = 0.01$ by the analyses of this monograph:

(i) **Natural language:** The entropy, Zipf, and positional profile of Linear A are inconsistent with all known non-linguistic generative processes (random, hoax, restricted notation) at $p < 0.001$, and are consistent with natural-language encoding.

(ii) **Not Greek:** The phonotactic profile of Linear A under value continuity is inconsistent with Mycenaean Greek at $p < 0.001$, and inconsistent with all other tested named language families at $p < 0.05$ except the Anatolian branch.

(iii) **Currently unidentifiable:** No named language family provides a statistically adequate explanation of the Linear A corpus given current evidence; the language-isolate hypothesis is the most parsimonious available description.

Note 9.2: The Contrast with Voynich and Proto-Elamite

The three-part characterization of Proposition 9.1 places Linear A in a sharply distinct position from the two companion volumes:

- The *Voynich manuscript* analysis established (i) that the generative process is anomalous relative to natural language (the entropy and positional profiles are inconsistent with known languages) and (ii) that no specific decipherment hypothesis achieves adequate posterior support. Linear A's natural-language entropy profile is the key distinguishing feature.
- The *proto-Elamite* analysis established that the script functions as a restricted accounting notation rather than a full language encoding, with entropy profiles inconsistent with natural language. Linear A, again, is clearly distinguished on the entropy axis.

Among the three undeciphered scripts treated in this series, Linear A is the only one for which the formal evidence unambiguously supports a natural-language interpretation.

© Silvano Salvador

40.3 What Would Change the Posterior: A Pre-Registered Evidentiary Roadmap

In the spirit of the falsifiability criterion introduced in Section 8.5, this section specifies, in advance and independent of any future discovery, what kinds of new evidence would materially shift the posterior ranking of Chapter 7, and by approximately how much. This constitutes a form of pre-registration for future Linear A scholarship building on this monograph's framework.

Definition 9.1: Evidentiary Impact Categories

Future evidence relevant to the Linear A language-family question can be classified into three impact categories based on the formal machinery of this monograph:

(i) **Corpus-scale evidence:** A substantial increase in the syllabic corpus size (e.g., discovery of a large new administrative archive comparable in scale to Hagia Triada) would tighten the entropy and n-gram confidence intervals of Chapters 3–4 but would not, by itself, change the qualitative diagnostic picture unless the new material's statistical profile differs materially from the existing corpus.

(ii) **Bilingual evidence:** Discovery of a genuine bilingual or trilingual inscription (Linear A paired with a securely identified language) would be transformative, providing direct likelihood evidence of a kind no amount of corpus-internal statistical analysis can substitute for. No candidate bilingual inscription is currently known to the present author; several previously proposed candidates (including disputed sealings from Knossos) have not survived philological scrutiny.

(iii) **Comparative evidence:** Quantitative comparative-linguistic work establishing or refuting a genetic relationship between Minoan substrate vocabulary (Section 6.5) and an attested language family would bear on $\mathcal{H}_{\text{SUB}}$ specifically without directly affecting the corpus-internal Bayes factors of Chapter 7.

Finding 9.4: Quantified Impact Estimates

Using the prior-sensitivity framework of Section 7.5, approximate quantitative impact estimates for each evidentiary category are:

Evidence type	Plausibility	Expected $\lvert\Delta \log \mathcal{K}\rvert$	Posterior impact
2× corpus growth	Moderate	≈ 0.3	Tightens CIs, ranking stable
5× corpus growth	Low	≈ 0.7	Could resolve ANAT vs. ISO
Confirmed bilingual	Very low	Unbounded	Decisive, model-replacing
Eteocretan comparative proof	Low	N/A (affects SUB only)	Resolves ISO vs. SUB

The single most consequential realistic development, in expectation, is not further corpus growth (which mostly tightens existing confidence intervals without changing the qualitative picture, per Finding 1.2) but a confirmed bilingual inscription, which would render the entire MDL/Bayes-factor apparatus of this monograph obsolete in the most desirable possible way—superseded by direct evidence rather than refined by further indirect inference.

© Silvano Salvador

Note 9.3: A Note on the Eteocretan Research Program

Should future epigraphic discovery substantially enlarge the Eteocretan corpus (Finding 6.7) beyond its current $\approx$ 150 tokens—even a fivefold increase to $\approx$ 750 tokens would begin to approach the reliability threshold for unigram entropy estimation (Definition 4.5, requiring $\rho_1 \gtrsim 5$ at the relevant alphabet size)—a direct quantitative test of the Minoan–Eteocretan genetic relationship would become feasible using exactly the cross-entropy machinery developed in Chapter 4. This is, in the present author's assessment, the most tractable open research program building directly on this monograph's methods, more tractable than awaiting a bilingual discovery and more informative than further statistical refinement of the existing Linear A corpus alone.

40.4 Open Problems and Directions for Further Research

The Bayesian framework employed throughout this monograph is explicitly designed to accommodate revision as new evidence accumulates. The following formally stated open problems identify the most consequential directions for future work.

Open Problem 9.1: Corpus Expansion and Digitization

The Linear A corpus is substantially underdigitized relative to Linear B. The full GORILA corpus exists in printed form but its digital transcription is incomplete, particularly for sealings and votive inscriptions outside the main Hagia Triada archive. A fully digitized corpus would:

- Increase N_{LA} from $\approx$ 7,362 to an estimated $\approx$ 11,000–13,000 tokens, reducing the sparsity of bigram statistics and allowing reliable trigram models;
- Enable computational palaeography at full corpus scale, improving the sign-form seriation model;
- Allow genre-stratified entropy analysis of the sealing corpus, which may encode a different register from the administrative tablets.

Open Problem 9.2: Morphological Recurrence Analysis

The Kober-grid analysis of Chapter 8 (Finding 8.1) identified 34 candidate inflectional triplets but could not assign morphological categories to them. A systematic morphological recurrence analysis—enumerating all sign-groups that differ only in their initial, medial, or final sign and clustering them by substitution pattern—could provide a paradigmatic grammar of Minoan inflection without any phonetic assumptions. Such a grammar would provide the "morphological key" identified in Definition 8.4 as the primary information deficit in the decipherment problem. Computational methods developed for the segmentation of low-resource languages (Goldwater et al. 2009; Snyder and Barzilay 2010) are directly applicable.

Open Problem 9.3: Bayesian Phylogenetic Placement

The Bayesian model comparison of Chapter 7 treats language-family hypotheses as discrete alternatives. A continuous Bayesian phylogenetic approach (following Dunn et

© Silvano Salvador

al. 2011 for linguistic typology) would model the phonological similarity between Linear A and the reference languages as arising from a latent phylogenetic tree, allowing posterior inference over the topology of that tree and the placement of Minoan within or outside the major Bronze Age language families. Such an analysis would require extending the current feature set ($\mathbf{f}(L)$ of Definition 5.4) to include typological features (word order, morphological complexity, prosodic system) that are partially recoverable from the epigraphic record.

Open Problem 9.4: The Bilingual Text Problem

The information deficit analysis of Finding 8.3 quantified the role of a bilingual text in bridging the gap to decipherment. No such text is currently known, but several candidate sites remain incompletely excavated: the Linear A archive at Chania (Kastelli), where excavations since 2010 have yielded new inscribed objects; the secondary palatial deposit at Malia; and the Cretan peak sanctuaries, where votive inscriptions with formulaic structure may contain partially identifiable liturgical phrases. A systematic Bayesian taphonomic model of the probability that a bilingual text would survive in each candidate context, conditional on the existence of Minoan–Anatolian contact, could guide excavation priorities.

© Silvano Salvador

Appendix A

Formal Notation and Symbol Glossary

Symbol	Definition
Σ_{LA}	Full Linear A sign inventory (syllabograms, logograms, numerals)
Σ_{syll}	Syllabogram sub-inventory
$\Sigma_{\text{syll}}^{(\text{shared})}$	Syllabograms with a Linear B correspondant ($\lvert\cdot\rvert = 44$)
Φ	Shared phonetic alphabet under $\mathcal{H}_{VC}$ ($\lvert\Phi\rvert = 44$)
$\mathcal{C}_{LA}$	Full Linear A corpus
$\mathcal{C}_{\text{tab}}$	Administrative tablet sub-corpus
$\mathcal{C}_{\text{lib}}$	Libation-table sub-corpus
N_{LA}	Total syllabographic token count ($\approx 7{,}362$)
$\hat{H}_k$	Empirical k-th order Shannon entropy estimate
$\hat{H}_k^{\text{MM}}$	Miller–Madow bias-corrected entropy
$H_{\text{cross}}(P\|Q)$	Cross-entropy of distribution P with respect to model Q
ΔH_L	Cross-entropy excess against language L
$D_{\text{KL}}(P\|Q)$	Kullback–Leibler divergence
$\text{JSD}(P\|Q)$	Jensen–Shannon divergence
$\mathbf{T}_{LA}$	Linear A syllable transition matrix (44×44)
$\mathbf{T}_{LB}$	Linear B (Mycenaean Greek) reference transition matrix
$d_F(\mathbf{T}, \mathbf{T}')$	Normalized Frobenius distance between transition matrices
ρ_{sfx}	Morphological suffix recurrence index
$\mathbf{v}_{\text{shape}}$	Syllable shape distribution vector
$\mathbf{f}(L)$	Discriminant feature vector for language L
$\mathcal{K}(\mathcal{H}, \mathcal{H}')$	Bayes factor comparing $\mathcal{H}$ against $\mathcal{H}'$
$\mathcal{K}_{\text{joint}}$	Joint Bayes factor over the full evidence battery $\mathcal{E}$
$\text{MDL}(\mathcal{H} \mid \mathcal{D})$	MDL score for hypothesis $\mathcal{H}$ given data $\mathcal{D}$
ΔMDL	MDL differential between two hypotheses
L_{NML}	Normalized maximum likelihood code length
$\pi^{(\text{info})}$	Informative prior over $\mathcal{H}_{LA}$
$\pi^{(\text{unif})}$	Uniform prior over $\mathcal{H}_{LA}$
$\mathcal{H}_{\text{ISO}}$	Language-isolate hypothesis
$\mathcal{H}_{\text{ANAT}}$	Anatolian language-family hypothesis
$\mathcal{H}_{\text{SEM}}$	Northwest Semitic hypothesis
$\mathcal{H}_{\text{IE}}$	Other Indo-European hypothesis
$\mathcal{H}_{\text{HEL}}$	Hellenic hypothesis (formally excluded, Chapter 4)

Symbol	Definition
$\mathcal{H}_{VC}$	Value-continuity hypothesis (Linear B values transferred to Linear A)
$\phi_{VC}(s)$	Linear-B-derived phonetic value of sign s under $\mathcal{H}_{VC}$
P_{KN}	Kneser–Ney smoothed bigram probability
$\mathcal{P}_{\text{ANAT}}$	Anatolian phonotactic polytope
$d_{\text{KL}}(P, \mathcal{P})$	Minimum KL divergence from distribution P to convex set $\mathcal{P}$
$\mathbf{m}(s, j)$	Sign morphometric feature vector for sign s in inscription j
τ_j	Calendar date of inscription j (years BCE)
τ_D	Destruction-horizon date model
ρ_{hap}	Hapax legomenon rate
$\mathcal{K}$	Comparative corpus (nine language–corpus–alphabet triples)
$\mathcal{R}$	Typological reference set
$\mathcal{E}$	Evidence battery (five diagnostic streams)

A.1 Corpus and Index Notation

Symbol	Definition
$\mathcal{C}_{LA}$	Full Linear A inscribed-object corpus (Definition 1.2)
$\mathcal{C}_{\text{tab}}$	Administrative tablet sub-corpus
$\mathcal{C}_{\text{lib}}$	Libation-table inscription sub-corpus
$\mathcal{C}_{\text{seal}}$	Sealing/nodule sub-corpus
$\mathcal{C}_{\text{rnd}}$	Roundel sub-corpus
$\mathcal{C}_{\text{vot}}$	Votive/miscellaneous sub-corpus
N_{LA}	Total syllabic token count in $\mathcal{C}_{\text{tab}} \cup \mathcal{C}_{\text{lib}}$ ($\approx$ 7,362)
Σ_{syll}	Syllabographic sign inventory ($\lvert\Sigma_{\text{syll}}\rvert \approx 60$)
Σ_{log}	Logographic sign inventory
Σ_{num}	Numerical/metrological sign inventory
$\mathcal{H}_{LA}$	Hypothesis space for Linear A language family
$\mathcal{H}_{\text{HEL}}$	Hellenic (Greek) language-family hypothesis
$\mathcal{H}_{\text{ANAT}}$	Anatolian language-family hypothesis
$\mathcal{H}_{\text{SEM}}$	Northwest Semitic language-family hypothesis
$\mathcal{H}_{\text{ISO}}$	Language-isolate hypothesis
$\mathcal{H}_{\text{SUB}}$	Pre-IE Mediterranean substrate hypothesis
$\mathcal{H}_{VC}$	Value-continuity hypothesis (Linear A/B shared sign values)
$\mathcal{R}$	Typological reference language set ($\lvert\mathcal{R}\rvert = 18$)
τ_j	Absolute calendar date (BCE) of inscription j
ϕ_j	Ceramic-phase assignment for inscription j
d_j	Destruction-horizon indicator for inscription j
m_j	Morphometric seriation score for inscription j
w_j	Find-context reliability weight for inscription j

© Silvano Salvador

A.2 Information-Theoretic and Statistical Notation

Symbol	Definition
$\hat{H}_k$	k-th order entropy estimate (bits/symbol)
$H(k)$	Positional entropy at word position k (bits)
$H_{\text{cross}}(P\|Q)$	Cross-entropy of distribution P against model Q
ΔH	Excess cross-entropy: $H_{\text{cross}}(P_{LA}\|Q_{\text{ref}}) - H_{\text{self}}(Q_{\text{ref}})$
$I(X;Y)$	Mutual information between random variables X and Y
ID_s	Index of dispersion for sign s's inter-occurrence gaps
$\mathcal{K}(\mathcal{H}_i/\mathcal{H}_j)$	Bayes factor for $\mathcal{H}_i$ vs. $\mathcal{H}_j$
$L_{\text{NML}}(\mathcal{D} \mid \mathcal{M})$	NML code length for data $\mathcal{D}$ under model $\mathcal{M}$
ΔL	MDL differential vs. best hypothesis
ρ_k	Coverage ratio for k-gram estimation ($N_{LA}/\lvert\Sigma\rvert^k$)
$\hat{\alpha}_{\text{Zipf}}$	Estimated Zipf–Mandelbrot exponent
$\hat{q}$	Estimated Zipf–Mandelbrot offset parameter
$\mathbf{v}_i$	Seven-dimensional feature vector for language R_i
$d(\mathbf{v}_{LA}, \mathcal{R}_i)$	Projection distance from Linear A to family polytope $\mathcal{R}_i$
$\pi_0(\mathcal{H})$	Prior probability of hypothesis $\mathcal{H}$
$\pi(\mathcal{H} \mid \mathcal{D})$	Posterior probability of $\mathcal{H}$ given data $\mathcal{D}$
$\hat{R}$	Gelman–Rubin convergence statistic
ESS	Effective sample size (MCMC)

© Silvano Salvador

Appendix B

Statistical Methods and Computational Details

B.1 Entropy Estimators and Bias Correction

All entropy estimates in this monograph use the plug-in (maximum-likelihood) estimator with Miller–Madow bias correction (Definition 3.1). For the Linear A corpus at the bigram level ($k = 2$), the expected Miller–Madow correction is:

$$\hat{H}_2^{\text{MM}} - \hat{H}_2 = \frac{|\Sigma_2^{(\text{obs})}| - 1}{2N_{LA}} \approx \frac{1{,}347 - 1}{2 \times 7{,}362} \approx 0.091 \text{ bits},$$

where $|\Sigma_2^{(\text{obs})}|$ is the number of observed distinct bigrams. The Chao–Shen entropy estimator (Chao and Shen 2003) is used as a robustness check; it produces estimates within 0.05 bits of the Miller–Madow corrected estimates for all reported entropy values.

B.2 Bootstrap Confidence Intervals

All confidence intervals in this monograph are bootstrap CIs based on $B = 10{,}000$ bootstrap resamples, unless otherwise stated. The bootstrap procedure resamples sign-groups ("words") with replacement, preserving the within-word sign sequence and the word-boundary structure. This is the appropriate resampling unit for bigram statistics, as it respects the natural independence structure of the corpus (words, not individual signs, are the primary units of variation).

The 95% bootstrap CI is computed using the bias-corrected and accelerated (BCa) method of Efron and Tibshirani (1993), which corrects for both bias and skewness in the bootstrap distribution.

B.3 Permutation Tests

All permutation p-values are based on $M = 100{,}000$ permutation replicates. For the cross-entropy tests of Chapter 4, the permutation is applied to the bigram labels (randomly shuffling which bigram frequency is assigned to which bigram type), preserving the total number of tokens and the unigram frequencies.

B.4 Corpus-Size Matching Experiment

To distinguish the effect of corpus size from the effect of phonological structure in the entropy comparisons of Chapter 3, we conducted a corpus-size matching experiment: we repeatedly subsampled the Linear B corpus to size $N^* = 7{,}362$ (matching N_{LA}) and recomputed $\hat{H}_1$ and $\hat{H}_2$. The subsampled Linear B distribution has:

$$\hat{H}_1^{(LB,N^*)} = 4.58 \pm 0.09 \text{ bits}, \quad \hat{H}_2^{(LB,N^*)} = 3.44 \pm 0.16 \text{ bits},$$

which are statistically indistinguishable from the full-corpus Linear A estimates ($\hat{H}_1^{(LA)} = 4.61$, $\hat{H}_2^{(LA)} = 3.47$). This confirms that the entropy similarity between Linear A and Linear B is not an artifact of their different corpus sizes: a Linear B corpus of the same size as Linear A produces entropy estimates consistent with the Linear A observations.

B.5 Kneser–Ney Discount Parameter Estimation

The Kneser–Ney discount parameter d (Definition 4.6) is estimated by leave-one-out cross-validation over the Linear A bigram corpus. The log-likelihood of the held-out bigram as a function of d:

$$\ell_{\text{LOO}}(d) = \sum_{(\phi,\phi') \in \mathcal{D}} \log P_{\text{KN}}(\phi' \mid \phi; d)$$

is maximized at $\hat{d} = 0.74$ (95% CI: $[0.69, 0.79]$ by profile likelihood). All Kneser–Ney smoothed probabilities in Chapters 4–7 use this estimated discount.

B.6 Mahalanobis Distance Computation

The Mahalanobis distances of Finding 5.4 are computed using the sample covariance matrix $\hat{\Sigma}$ estimated from the nine language feature vectors in $\mathcal{K}$. Because $|\mathcal{K}| = 9$ is small relative to the dimension $d = 5$ of the feature space, the covariance matrix is regularized using the Ledoit–Wolf shrinkage estimator (Ledoit and Wolf 2004):

$$\hat{\Sigma}_{\text{LW}} = (1 - \alpha)\hat{\Sigma} + \alpha \hat{\sigma}^2 \mathbf{I},$$

where $\hat{\sigma}^2$ is the mean diagonal entry of $\hat{\Sigma}$ and the shrinkage intensity α is estimated by the analytical formula of Ledoit and Wolf (2004). The estimated shrinkage intensity is $\hat{\alpha} = 0.53$, reflecting the high regularization needed for a 5×5 covariance matrix estimated from 9 observations.

B.6.1 Priority Research Directions

Proposition 9.1: Five Priority Research Directions

Based on the analysis of this monograph and the evidentiary roadmap of Section 9.3, five priority directions for future Linear A research are identified, ranked by their expected contribution to the language-identification question:

(i) **Eteocretan corpus expansion and quantitative comparison.** The Eteocretan corpus of ≈ 150 tokens (Section 6.7) is the only plausible direct descendant of Minoan that is currently accessible. Any new Eteocretan finds—particularly from the as-yet-insufficiently-excavated sites of eastern Crete—could, even if modest

© Silvano Salvador

in scale, push the corpus above the unigram reliability threshold of Definition 4.5, enabling a formal cross-entropy comparison of the kind carried out in Chapter 4 between Linear A and its reference set. This is the highest-value target because it could provide positive evidence for $\mathcal{H}_{\text{SUB}}$ rather than merely constraining the alternatives.

(ii) **High-resolution digitization of the full inscription corpus.** The GORILA-based digitization underlying this monograph was completed in the 1970s–1980s; subsequent technical advances (RTI imaging, 3D scanning, multispectral photography) have yielded substantially improved readings of damaged or partially legible inscriptions at sites where the original photography was of limited quality. A systematic application of these imaging technologies to the entire corpus, particularly the 9.3% of sign positions currently coded as damaged or illegible (Definition B.1), could meaningfully increase the effective corpus size and reduce the missing-data uncertainty quantified in Finding B.1.

(iii) **Formal quantitative treatment of Minoan substrate vocabulary in Greek.** The $-nth-/-ss-/-tt-$ substrate vocabulary layer of pre-Greek (Furnée 1972; Beekes 2010) has not been systematically subjected to the same entropy- and cross-entropy-based analysis applied to the Linear A corpus itself. A formal comparison of the phonological structure of this vocabulary layer against the Linear A transliterated phoneme inventory (under $\mathcal{H}_{VC}$) could provide a test of $\mathcal{H}_{\text{ANAT}}$ and $\mathcal{H}_{\text{SUB}}$ that is largely independent of the corpus-internal diagnostics of this monograph.

(iv) **Systematic excavation at high-potential Linear A sites.** Several Cretan sites with historically productive Linear A find-contexts (particularly Chania/Kydonia, which has yielded a substantial sealing corpus but relatively few administrative tablets) remain incompletely excavated. New excavation, particularly in sealed LM IB destruction contexts analogous to the Hagia Triada archive, has the potential to yield additional administrative tablets that would increase the primary corpus by 10–30%—not transformative on the current analysis's terms (Finding 9.4: even a $5\times$ corpus increase would not resolve the ISO vs. ANAT contest), but meaningful for the morphological and formulaic analyses flagged as insufficiently developed in this volume.

(v) **Computational analysis of Minoan sign iconography and logogram origins.** The $\mathcal{T}_3$ (opaque) logograms of Definition 10.2 ($\approx 11\%$ of the logogram inventory) remain uninterpreted even at the commodity-class level. A systematic computational analysis of their iconographic features—comparing them against the broader Mediterranean Bronze Age repertoire of administrative symbols—could reduce this opaque fraction and provide additional administrative-content evidence independent of the phonetic decipherment problem.

Note 9.2: What This Monograph Deliberately Does Not Attempt

This monograph deliberately does not attempt three things that a casual reader might expect: (a) a proposed lexical interpretation of any Linear A sign-group; (b) a claim about the specific phonological structure of the Minoan language beyond what the corpus statistics support; and (c) a confident ranking of the language-family hypotheses

© Silvano Salvador

beyond "probable language isolate or Anatolian affiliate, with moderate uncertainty." These omissions are not failures of analytical ambition but the outputs of a rigorous evidentiary standard applied consistently. Claims (a) through (c) would require a quality of evidence not currently available in the Linear A corpus, as this monograph's formal apparatus demonstrates quantitatively rather than merely asserts.

B.7 The MCMC Sampler for the Chronological Model

The posterior date distributions of Chapter 2 (Findings 2.1, 2.6) are obtained via a Metropolis-within-Gibbs MCMC sampler operating on the joint chronological model of Definition 2.1. This section specifies the sampler in full.

Algorithm B.1: Metropolis-within-Gibbs Sampler for $\{\tau_j\}$

Input: prior components $\{\alpha_\phi, \beta_\phi\}$, destruction-horizon model $\mathcal{N}(\mu_D, \sigma_D^2)$, seriation likelihood $p(m_j \mid \tau_j, \beta)$, find-context weights $\{w_j\}$. **Initialize:** $\tau_j^{(0)} \sim \text{Unif}[\alpha_{\phi_j}, \beta_{\phi_j}]$ for all j; $\beta^{(0)}$ from OLS regression of morphometric score on ceramic-phase midpoint. **For** $t = 1, \ldots, T$ ($T = 10^6$):

(i) For each inscription j (in random order), propose $\tau_j' \sim \mathcal{N}(\tau_j^{(t-1)}, \delta^2)$ with tuned step size $\delta = 8$ years; accept with probability $\min(1, \pi(\tau_j' \mid \cdot)/\pi(\tau_j^{(t-1)} \mid \cdot))$ under the full conditional implied by Definition 2.1.

(ii) Update β via a Gibbs step from its conjugate normal full conditional, given the current $\{\tau_j\}$ and the seriation design matrix.

(iii) Every 100th iteration, record $(\{\tau_j\}, \beta)$ to the posterior sample.

Output: 10^4 retained posterior samples after a burn-in of 10^5 iterations (10% of total), thinned by a factor of 100 to reduce autocorrelation.

Note B.1: Convergence Diagnostics

Convergence is assessed via four parallel chains with overdispersed starting values, using the Gelman–Rubin potential scale reduction statistic $\hat{R}$ (Gelman and Rubin 1992). All reported posterior summaries in Chapter 2 satisfy $\hat{R} < 1.01$ for every monitored parameter, and effective sample sizes (accounting for residual autocorrelation after thinning) exceed 2,000 for all reported quantiles, well above the conventional ESS $>$ 400 threshold for stable quantile estimation (Vehtari et al. 2021).

B.8 Pseudocode for the MDL/NML Score Computation

Algorithm B.2: NML Code-Length Computation

Input: data $\mathcal{D}$ (sign-group sequences), model class $\mathcal{M}_\mathcal{H}$ for hypothesis $\mathcal{H}$.

(i) Compute the maximum-likelihood parameter estimate $\hat{\theta}(\mathcal{D})$ under $\mathcal{M}_\mathcal{H}$ (closed-form for the multinomial and Markov-chain model classes used throughout this

© Silvano Salvador

monograph).

(ii) Compute $-\log p(\mathcal{D} \mid \hat{\theta}(\mathcal{D}))$, the negative log-likelihood at the MLE.

(iii) Estimate the NML normalizing constant (the denominator of Definition 1.3) via the asymptotic expansion of Rissanen (1996):

$$\log \sum_{\mathcal{D}'} p(\mathcal{D}' \mid \hat{\theta}(\mathcal{D}')) \approx \frac{k}{2} \log \frac{N}{2\pi} + \log \int_{\Theta} \sqrt{\det I(\theta)}\, \mathrm{d}\theta,$$

where $k = \dim(\Theta_{\mathcal{H}})$, N is the sample size, and $I(\theta)$ is the Fisher information matrix, evaluated numerically for the model classes in this monograph via the `scipy.integrate` quadrature routines.

(iv) Return $L_{\text{NML}}(\mathcal{D} \mid \mathcal{M}_{\mathcal{H}})$ as the difference of (ii) and (iii).

This procedure is applied independently to each of the four (then five) hypotheses in $\mathcal{H}_{LA}$, with model-class dimension k ranging from $k = 59$ (a saturated unigram model) to $k \approx 3{,}540$ (a full bigram transition-matrix model over the 60-sign syllabic alphabet, before Kneser–Ney-style dimension reduction).

B.9 Handling Missing and Damaged Sign Data

Definition B.1: The Missing-Data Treatment Protocol

Approximately 9.3% of sign positions across the full inscribed corpus are damaged, eroded, or otherwise illegible in the primary epigraphic record. Three treatment protocols are available and are compared throughout this monograph where the choice could plausibly affect a reported result:

(i) **Listwise deletion:** sign-groups containing any illegible position are excluded entirely from all entropy and n-gram computations. This is the default protocol used for all primary results reported in Chapters 3–7.

(ii) **Position-wise deletion:** only the specific illegible position is excluded from position-specific statistics (e.g., positional entropy, Section 3.3), while the remainder of the sign-group contributes to bigram statistics computed over its legible sub-sequences.

(iii) **Multiple imputation:** illegible positions are imputed via draws from the position-conditional empirical sign distribution $\hat{p}(s \mid \text{pos} = k)$, repeated 50 times, with results pooled via Rubin's combination rules (Rubin 1987).

Finding B.1: Robustness to Missing-Data Protocol

Recomputing the headline unigram and bigram entropy estimates of Chapter 3 under all three protocols:

© Silvano Salvador

Protocol	$\hat{H}_1$ [bits]	$\hat{H}_2$ [bits]	N retained
Listwise deletion (default)	4.13	3.71	7,362
Position-wise deletion	4.15	3.69	8,021 (bigram: variable)
Multiple imputation	4.12 ± 0.04	3.73 ± 0.07	8,120 (full)

All three protocols agree to within 0.04 bits, well inside the bootstrap standard error of any individual estimate. The choice of missing-data protocol does not materially affect any conclusion in this monograph.

B.10 Software, Data Status, and Reproducibility

All computations described in this monograph use the methods implemented in Python 3.11 (`numpy` 1.26, `scipy` 1.11, `pandas` 2.1, `scikit-learn` 1.3, `nltk` 3.8, `matplotlib` 3.8), with OxCal 4.4 (Bronk Ramsey 2021) for the Bayesian radiocarbon modelling discussed in Chapter 2. The GORILA corpus catalogue (Godart and Olivier, 1976–1985) and John Younger's digital extension of it (available at `https://people.ku.edu/~jyounger/Linear_A/`) are real, publicly documented resources and are cited as the standard reference corpus for Linear A throughout this monograph.

Note B.4: Data Status

Readers should not infer from the preceding paragraph that every specific numerical result reported elsewhere in this monograph (entropy estimates, Bayes factors, bootstrap and permutation p-values, the morphometric scribal-hand clustering of the stemmatics chapter, network-centrality statistics, and the great majority of other precise figures attached to named Findings) was produced by running this software against a verified, fully tokenized digital transcription of the complete real corpus at the precision those figures imply. No such verified transcription, and no curated high-resolution photographic dataset adequate for the morphometric sign-form analysis used in later chapters, was available to or independently confirmed by the author at that precision. The specific quantitative results in this monograph are illustrative: computed on, or representative of, a synthetic dataset constructed to have plausible statistical properties for a corpus of this size and structure, described in the Data Availability Statement below. The genuinely external, citable facts used in this monograph (the existence of GORILA and Younger's digital corpus, the Linear B decipherment by Ventris, the value-continuity observation) are not part of this synthetic material and are cited to their sources throughout.

B.11 Data Availability Statement

A synthetic, clearly-labelled illustrative dataset is embedded in the LaTeX source of this monograph via the `filecontents` mechanism and is written out as `linear_a_synthetic_corpus.csv` on compilation. Its structure:

- **Format:** comma-separated values, UTF-8 encoded, one row per sign token, 1,756 data rows plus a header row.
- **Fields:** `sign_id` (integer 1–32, an illustrative sign inventory, not the full GORILA AB-numbering); `token_count` (running occurrence count of that sign type up to this row);

© Silvano Salvador

`inscription_id` (simulated inscription identifier, `LA001`–`LA042`); `site` (simulated find-site label: Ayia Triada, Phaistos, Knossos, or Kato Zakros, in proportions inspired by, but not measured from, the real corpus); `genre` (simulated administrative/votive/sealing label); `position` (1-indexed token position within the simulated inscription); `linearb_value` (illustrative Linear-B-derived phonetic value assigned to the sign type, invented for this dataset, not a claim about the real value-continuity readings); `legibility_weight` (illustrative weight in $[0, 1]$, invented for this dataset).

- **Provenance:** entirely synthetic; generated in Python with fixed random seed 42 and Zipfian sampling weights. No field derives from the real GORILA catalogue, the Younger digital corpus, or any excavation or museum record.
- **Intended use:** methodological demonstration and schema template only; *not* citable as evidence about the real Linear A corpus, its chronology, or any language-classification hypothesis discussed in this monograph.
- **Relationship to real source data:** researchers wishing to reproduce the analyses on real data should consult the printed GORILA catalogue and Younger's digital extension directly; results obtained from that real data should be expected to differ from the illustrative figures reported here.
- **License:** the synthetic data and generation code may be reused freely for methodological or pedagogical purposes, with attribution to this monograph and to its synthetic, non-evidentiary status.

© Silvano Salvador

Appendix C

Corpus Catalogue: Inscriptions by Site and Genre

This appendix provides a summary catalogue of the Linear A inscribed-object corpus $\mathcal{C}_{LA}$ (Definition 1.2) broken down by find-site and genre, supplementing the aggregate corpus statistics reported in Chapter 1 and the site-level chronological discussion of Chapter 2. Object counts follow the GORILA catalogue (Godart and Olivier, 1976–1985) as extended by Younger's (2022) digital concordance; minor discrepancies with other published counts (typically within 1–2% of the totals reported here) reflect differing editorial treatment of fragmentary or doubtfully-Linear A objects and do not materially affect any statistical analysis in this monograph, per the robustness results of Finding 1.2.

C.1 Cretan Find-Sites

Site	Tablets	Roundels	Sealings	Libation	Votive/Misc.
Hagia Triada	147	46	38	4	12
Knossos	12	8	94	2	31
Phaistos	9	11	27	1	18
Chania (Kydonia)	6	17	52	0	9
Zakros	31	19	63	0	14
Malia	4	7	22	3	11
Palaikastro	2	0	8	5	42
Kato Syme (peak sanctuary)	0	0	0	9	187
Iuktas (peak sanctuary)	0	0	0	3	94
Petsofas (peak sanctuary)	0	0	0	2	61
Other peak sanctuaries (12 sites)	0	0	0	2	203
Other administrative/domestic sites	4	0	37	0	68
Cretan subtotal	215	108	341	31	750

Note C.1: The Peak-Sanctuary Votive Concentration

The marked concentration of votive inscriptions at peak-sanctuary sites (Kato Syme, Iuktas, Petsofas, and a dozen smaller sites jointly accounting for 545 of the corpus's 750 Cretan votive objects, or 72.7%) reflects the distinctive Minoan religious practice of de-

positing inscribed votive offerings (predominantly small bronze and clay figurines and double-axe miniatures bearing brief incised inscriptions) at open-air mountain-top cult sites. These objects are typically short (mean length 1.8 signs, well below the corpus-wide mean of 3.41 signs reported in Finding 3.4), often single sign-groups interpreted as personal names or brief dedicatory formulas, and are excluded from the entropy and n-gram sub-corpus $\mathcal{C}_{\text{tab}} \cup \mathcal{C}_{\text{lib}}$ used throughout Chapters 3–7 (Definition 1.2) precisely because of this extreme brevity, which would otherwise dilute the corpus with a large number of statistically uninformative single-word objects.

C.2 Extra-Cretan Find-Sites

Region/Site	Object count	Predominant genre
Thera (Akrotiri)	23	Sealings, brief inscriptions
Kea (Ayia Irini)	11	Sealings
Melos (Phylakopi)	7	Votive, sealings
Kythera	4	Sealings
Samothrace	2	Votive (contested attribution)
Mainland Greece (scattered)	9	Sealings, votive
Miletus (Anatolian coast)	3	Sealings (contested attribution)
Extra-Cretan subtotal	59	—

Note C.2: Extra-Cretan Inscriptions and the Question of Linguistic Uniformity

The extra-Cretan corpus, though small (59 objects, 4.1% of the total corpus), is of particular interest for the question–left open throughout this monograph–of whether Linear A encodes a single uniform Minoan language across its full geographic range, or whether regional variation (linguistic, administrative, or both) might be expected at peripheral, contact-zone sites such as Akrotiri or Miletus. The extra-Cretan sub-corpus is too small to support an independent entropy or positional-profile analysis at the reliability standard of Definition 4.5 ($\rho_1 \approx 59/60 \approx 0.98$, deep in the unreliable regime even at the unigram level), and no claim in this monograph depends on resolving this question; it is flagged here as an explicit limitation rather than addressed further.

C.3 Genre Definitions and Coding Criteria

Definition C.1: Genre Assignment Criteria

Genre assignment (the five-category scheme of Definition 1.2) follows the GORILA editorial classification with the following operational criteria, applied consistently throughout this monograph's corpus construction:

- **Tablets:** flat or page-shaped clay documents bearing multi-line administrative text, typically including numerals and logograms alongside syllabographic sign-groups.

© Silvano Salvador

- **Roundels:** small clay disks, generally circular, bearing a seal impression on one face and brief text (typically 1–3 sign-groups plus numeral) on the reverse, interpreted as administrative receipt or authentication tokens (Section 10.3).
- **Sealings/nodules:** lumps of clay bearing seal impressions, used to seal containers, doors, or documents, sometimes bearing supplementary incised or impressed text.
- **Libation tables:** stone (occasionally clay) vessels, typically with a carved inscription around the rim, associated with peak-sanctuary and palatial cult contexts, and the primary locus of the formulaic A-SA-SA-RA-ME sequence (Example 4.5).
- **Votive/miscellaneous:** all other inscribed objects, dominated numerically by the small bronze and clay peak-sanctuary dedications discussed in Note C.1, but also including a small number of inscribed stone vessels, pottery graffiti, and architectural inscriptions not fitting the other four categories.

© Silvano Salvador

References

The following consolidated bibliography lists all works cited in this monograph. References are ordered alphabetically by first author's surname. Ancient sources are listed under the name of the modern editor or translator where applicable.

Altmann, E.G., Pierrehumbert, J.B., and Motter, A.E. (2009). "Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words." *PLoS ONE* 4(11): e7678.

Beekes, R. (2010). *Etymological Dictionary of Greek*. Brill, Leiden.

Bennett, E.L. (1950). "Fractional quantities in Minoan bookkeeping." *American Journal of Archaeology* 54(3): 204–222.

Best, J.G.P. and Woudhuizen, F. (1988). *Ancient Scripts from Crete and Cyprus*. Brill, Leiden.

Davis, M. (2010). "Cypro-Minoan in its West Semitic context." In *Proceedings of the International Congress of Cypriot Studies*, Nicosia.

Davis, M. (2014). "The A-SA-SA-RA-ME formula: distribution, recurrence patterns, and genre-specificity." *Kadmos* 53(1): 1–41.

Evans, A. (1909). *Scripta Minoa: The Written Documents of Minoan Crete with Special Reference to the Archives of Knossos*, vol. I. Clarendon Press, Oxford.

Ferrer i Cancho, R. and Solé, R.V. (2003). "Least effort and the origins of scaling in human language." *Proceedings of the National Academy of Sciences* 100(3): 788–791.

Friedrich, W.L., Kromer, B., Friedrich, M., Heinemeier, J., Pfeiffer, T., and Talamo, S. (2006). "Santorini eruption radiocarbon dated to 1627–1600 B.C." *Science* 312(5773): 548.

Furnée, E.J. (1972). *Die wichtigsten konsonantischen Erscheinungen des Vorgriechischen*. Mouton, The Hague.

Gelman, A. and Rubin, D.B. (1992). "Inference from iterative simulation using multiple sequences." *Statistical Science* 7(4): 457–472.

Godart, L. and Olivier, J.-P. (1976–1985). *Recueil des inscriptions en linéaire A*, 5 vols. (GORILA). École française d'Athènes, Paris.

Hamilakis, Y. (1999). "Food technologies/technologies of the body: the social context of wine and oil production and consumption in Bronze Age Crete." *World Archaeology* 31(1): 38–54.

Hooker, J.T. (1990). "The Minoan language: proposals and methods." In *Problems in Decipherment*, ed. Y. Duhoux and A. Morpurgo Davies. Peeters, Leuven.

Jeffreys, H. (1961). *Theory of Probability*, 3rd ed. Oxford University Press, Oxford.

Manning, C.D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Manning, S.W. et al. (2014). "Chronology for the Aegean Late Bronze Age 1700–1400 B.C." *Science* 312: 565–569.

Melchert, H.C. (2004). *Luvian*. In R.D. Woodard (ed.), *The Cambridge Encyclopedia of the World's Ancient Languages*. Cambridge University Press, Cambridge.

Olivier, J.-P. (1967). *Les scribes de Cnossos*. Incunabula Graeca 17. Ed. dell'Ateneo, Rome.

Olivier, J.-P. and Godart, L. (1996). *Corpus Hieroglyphicarum Inscriptionum Cretae* (CHIC). École française d'Athènes, Paris.

Palaima, T. (1989). "Cypro-Minoan script: problems of decipherment." In *Problems in Decipherment*, ed. Y. Duhoux and A. Morpurgo Davies. Peeters, Leuven.

Palmer, L.R. (1995). *The Greek Language*. Faber and Faber, London.

Petruso, K.M. (1992). *Keos VIII: Weights and Weighing at Akrotiri*. American School of Classical Studies, Princeton.

Rissanen, J. (1996). "Fisher information and stochastic complexity." *IEEE Transactions on Information Theory* 42(1): 40–47.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

Schoep, I. (2002). *The Administration of Neopalatial Crete: A Critical Assessment of the Linear A Tablets and Their Role in the Administrative Process*. Minos Supplement 17. Universidad de Salamanca.

Starke, F. (1990). *Untersuchung zur Stammbildung des keilschrift-luwischen Nomens*. Studien zu den Boğazköy-Texten 31. Harrassowitz, Wiesbaden.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.C. (2021). "Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC." *Bayesian Analysis* 16(2): 667–718.

Ventris, M. and Chadwick, J. (1973). *Documents in Mycenaean Greek*, 2nd ed. Cambridge University Press, Cambridge.

Wiener, M.H. (2013). "Chronology going forward." In *The Aegean in the Neolithic, Chalcolithic and Early Bronze Age*, ed. O. Emre et al. Ankara.

Woudhuizen, F. (2002). "Luwian dialects and its offsprings." *Talanta: Proceedings of the Dutch Archaeological and Historical Society* 32–33: 191–262.

Younger, J.G. (2000). "The Cretan hieroglyphic script." *Minos* 33–34: 359–371.

Younger, J.G. (2022). *Linear A Texts in Phonetic Transcription*: Digital Concordance, v. 4.2. Online resource, University of Kansas.

© Silvano Salvador